ANALYSIS OF MULTIVARIATE RESPONSES IN PATIENT REPORTED OUTCOME MEASURES: MISSING DATA AND AUXILIARY VARIABLES

By

Olawale F. Ayilara

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfillment of the requirements of the degree of

Doctor of Philosophy

Department of Community Health Sciences

Rady Faculty of Health Sciences

University of Manitoba

Winnipeg

ABSTRACT

Patient-reported outcomes measures (PROMs) are increasingly used in clinical registries, cohort studies, population-based surveys and clinical trials to collect information about patient's perspectives of their own health, including physical, mental, and social health. PROMs instruments are multivariate (i.e., multiple correlated items or questions). Item non-response or missing data, which may occur when patients fail to complete or respond to PROMs question, threatens the validity of findings from the assessment of group differences or longitudinal change in PROMs. The goal of this research was to develop and evaluate methods for addressing item non-response in PROMs. Four related studies were undertaken.

The first study compared the performance of non-negative matrix factorization (NNMF), which is an unsupervised machine-learning method that uses optimization techniques to detect a low-dimensional structure from the data, with full information maximum likelihood (FIML) and listwise deletion. The methods were applied to test for differential item functioning in multidimensional PROMs using Monte Carlo simulation and real-world data. We found that the Type I error rates and statistical power for the NNMF method were comparable to the FIML method. However, the NNMF method performed less than optimal when the correlations amongst items were weak.

The second study evaluated the performance of NNMF, FIML, multiple imputation (MI) with conditional proportional odds model, and listwise deletion methods when estimating longitudinal change in latent variable means. Simulated and real-world data were used to compare the methods. We found that the NNMF is relatively efficient when sample size is large and the percentage of non-response is high, but less optimal under other data-analytic conditions.

The third study investigated the use of auxiliary or supplementary variables in imputation model and compared the precision and bias of FIML, MI with and without auxiliary variable, when estimating longitudinal change in PROM scores. We found that including auxiliary information in the imputation model increased the precision and reduced the bias of the estimated parameters.

The fourth study proposed an enhanced weighted NNMF, which uses observed item responses as auxiliary variable to define weights for item level imputation, and compares the performance with FIML, NNMF and listwise deletion. Similar to other studies, we used simulated and real-world data to compare these methods. The proposed method outperformed the NNMF method for most of the simulation conditions under consideration, especially when the number of items was large. The proposed method resulted in inflated Type I error rates when the percentage of missing responses was more than 25% and the number of items was small.

This research contributes to the statistical literature on methods to address missing data in PROMs with potential applications in clinical and quality of life research. Also, it demonstrates the practicality of using observed item responses to define an auxiliary variable, which provides a basis for easy, accessible, and cost-effective approach of identifying auxiliary variable in PROMs. Finally, the research findings benefit psychometricians and other researchers who use PROMs to understand the complex relationship that exist among ordinal items and underlying latent variables.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisors, Drs. Lisa Lix and Mohammad Jafari Jozani, for all their support throughout my PhD program. Their timely feedback and contributions helped me while conducting this research and writing the manuscripts. I would like to thank members of my advisory committee – Drs. Tolulope Sajobi and Ruth Barclay for their suggestions, and constructive criticisms of this research. I would also like to thank Drs. Robert Platt and Pourang Irani for allowing me to do my summer internship with their research groups. Thanks to Drs. Eric Bohm and Richard Sawatzky for providing the clinical data used in this research.

I am immensely grateful to all the member of my family for their support and understanding throughout my graduate studies. Special thanks to my wife, Ifeoluwa Ayilara, for her enormous support, sacrifices and words of encouragement. I could not have made it through my PhD program without you.

My graduate studies dream in Canada would not have been possible without the support of Engr. Akin Odumakinde, who sponsored all my initial cost of settling in Canada including airfare. I would like to thank Mr. Joseph Baiyekusi (Impact Your World Leadership Initiative) and Mrs. Tina Aizebeokhai for their support and guidance during my visa application. I am grateful to my friends, Erfan Hoque, Faisal Atakora, Yixiu Liu and all the members of Dr. Lix's research group for making my PhD experience remarkable.

My PhD program was funded by the Visual and Automated Disease Analytic Program Scholarship, the Canadian Institute of Health Research (Institute Community Support Program Travel Awards) and the University of Manitoba International Graduate Student Entrance Scholarship. The Compute Canada (now, Digital Research Alliance of Canada) provided high-

performance computers to run my simulations. Centre for Healthcare Innovation provided me with office space to conduct this research.

DEDICATION

This dissertation is dedicated to the following people:

Late Mrs. Olufunmilayo Ayilara, my mother, who sacrificed everything so that I could have the best education. I am eternally grateful to you mummy.

Ifeoluwa Ayilara, my wife, who supported and motivated me throughout my PhD program. Thank you for the pep talk and for sharing in my pains and joy as a graduate student. I could not have made it through without you; I love you IfeWales.

Oluwasemilore (Elizabeth) Ayilara and Ireoluwa (Naomi) Ayilara, my beautiful daughters, who were born during my PhD program. Your arrivals have brought happiness, goodness and joy to our family, I love you Lore and Ire.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Overview

Patient-reported outcome measures (PROMs) are used in clinical registries, cohort studies, population-based surveys and clinical trials to measure health-related quality of life (HRQOL) [1–3]. HRQOL is a self-perceived, multidimensional construct, which includes physical, psychological, and social functioning, that are influenced by disease and/or various forms of treatment [4, 5]. PROMs are used to describe disease trajectories [6], determine and evaluate the impact of interventions, such as therapy [7]. PROMs can complement other sources of data that contain information about an individual's health, such as clinical and administrative health data, to inform assessments of quality of care and effectiveness of health policies and services [3, 8, 9].

A PROMs instrument can be generic (i.e., it can be used in different groups or populations), or disease-specific (i.e., it can be used to measure outcomes that are characteristic of a specific disease). The 12-item Short Form Health Survey (SF-12), or a longer version, the SF-36 [10], are generic instruments that aim to capture information about physical and mental health [9, 11]. Examples of disease-specific instruments include the Oxford Hip and Knee instruments, which are 12-item questionnaires designed to assess pain and function of the hip and knee, respectively [12, 13]. Response options for the items or questions on a PROMs instrument are often ordinal (e.g., Likert scales).

The quality and value of findings from PROMs are dependent on a number of factors. Some of the key factors include: (1) accuracy of mapping the items to a latent (i.e., unobserved) construct or variable they are designed to measure [14], (2) extent to which the PROMs instrument

measures a latent variable with sufficient consistency (i.e., extent to which the same results would be obtained repeatedly) [15], and (3) extent to which the instrument measures what it is designed to measure [16].

Statistical analysis of PROMs data relies on the assumption that patients or individuals interpret and respond to questions about their health in the same way, such that scores are equally applicable to all patients in the population [17–20]. This assumption stems from the invariance property of the PROMs instrument. The invariance property, which asserts that the measuring instrument is independent of what it is measuring, is required to make valid inferences and reliable comparisons of PROM scores across groups or time [20–22]. This assumption may not always be tenable, especially in a heterogeneous population. For example, patients with the same underlying level of health may respond to PROM questions in systematically unique ways because of cultural, developmental or personality differences, contextual factors or life circumstances, and different health experiences. This is known as differential item functioning (DIF); if unaccounted, DIF can result in an unexpected lack of scale comparability, and erroneous conclusions about the presence of group differences [23–27].

PROMs instruments are often multivariate (i.e., multiple correlated items or questions); the data arising from the instruments can be analyzed at the item-level, subscale or scale-level. Methods for analyzing PROMs data can be broadly divided into: (1) classical test theory (CTT) or true-score theory, and (2) latent variable models (LVMs). CTT assumes that the latent variable is continuous. The method of estimation is based on the true-score model, which relates individual's observed score (i.e., unweighted sum of the individual's responses to items) to his or her location on the latent variable [28]. CTT assumes that each observed score is a linear combination of a true score and random error. LVMs such as exploratory or confirmatory factor

analysis, and item response theory also assumes that the latent variable is continuous. LVMs relates a set of items to one or more latent variables. In contrast to the CTT in which the individual's observed score is the unit of focus, in LVMs the item or subscale is the unit of focus.

Most studies that use PROMs are conducted at the subscale level or the scale level without a detailed evaluation of the individual item-level data; this may be due, in part, to the limited set of analytic tools for ordinal data [29, 30], particularly for the treatment of missing data. Missing data are defined as "values that are not available and that would be meaningful for analysis if they were observed" [31]. Missing data can be broadly classified as item non-response or unit non-response. Item non-response occurs when a patient responds to some but not all the items because they are too sick, or have difficulty understanding or interpreting the items, or uncomfortable responding to the items. Unit non-response occurs when an individual does not respond to the items of an instrument because they cannot be reached for interview or miss all scheduled clinic visits. Methods discussed in this thesis focus on item non-response.

The potential consequences of item non-response are: (i) loss of statistical power, which may increase the risk of false-negative findings, and (ii) over- or under-estimation of differences between groups or over time [32, 33]. The choice of methods to address item non-response is dependent, in part, on the missingness mechanism, which includes missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR) [34]. Data are MCAR when the probability of non-response on an item is unrelated to responses of other items and the item itself. Data are MAR when the probability of non-response on an item is associated with responses on some other items but not the response of the item itself. Finally, data are MNAR when the probability of non-response on an item is related to the responses of the item itself, and other items.

Statistical methods to address item non-response can be broadly divided into the following: (1) deletion methods, (2) imputation methods, and (3) likelihood-based methods. Deletion methods such as listwise deletion or complete-case analysis, and pairwise deletion or available-case analysis are commonly used to address item non-response because they can be easily implemented and are available in many statistical packages. The listwise deletion method removes individuals with at least one item with a missing response. The pairwise deletion method removes individuals on an analysis-by-analysis basis depending on the available responses. Both methods assume data are MCAR and can produce bias parameter estimates when this assumption is not plausible. Deletion methods are potentially wasteful and can reduce statistical power.

Imputation refers to replacing a missing value by a plausible value or a set of plausible values. Imputation methods can be broadly categorized as single and multiple imputation. Single imputation methods, such as unconditional mean or mean imputation [32], and conditional mean imputation or regression imputation [34], generate a single plausible value for each missing value. Single imputation methods seem advantageous over deletion methods because they reduce or eliminates data wastage. Regardless of the advantage, these methods underestimate the variability in the data because the plausible values generated are treated as true values. Hence, these methods ignore the uncertainty due to the imputation. Other single imputation methods, such as stochastic regression imputation [32], eliminate the biases associated with unconditional and conditional mean imputation by incorporating some randomness to the generated plausible values. Stochastic regression imputation produces unbiased estimates when data are MCAR or MAR. However, like the unconditional and conditional mean imputation methods, stochastic

regression imputation reduces standard errors of the parameters, resulting in an increased potential for false-positive results [35].

Multiple imputation (MI) methods such as multivariate normal imputation [36] and fully conditional specification or chained equations or sequential regression imputation [37, 38], replaces missing value with different sets of plausible values. These methods are widely used to address MCAR or MAR data. The multivariate normal imputation method assumes that all variables in the imputation model jointly follow a multivariate normal distribution, which are often violated when variables are ordinal [36, 39].

The fully conditional specification method is more flexible and does not rely on the assumption of multivariate normality. This method replaces the missing values on a variable-by-variable basis, with the matching distribution of the variable in the imputation model [40, 41]. For example, a linear regression model is used to impute continuous variables; logistic regression model to impute binary variables; and proportional odds model to impute ordinal variables [42]. Multiple imputation methods require that the functional form of the imputation model and the distributional form of the variables included in the model are specified correctly [42, 43]. Also, potential non-linear relationships that may exist amongst variables needs to be incorporated in the imputation model [44].

Likelihood-based methods, such as direct likelihood or full information maximum likelihood (FIML), rely on the log-likelihood function to choose parameter values that are most likely to have produced the observed data [34, 35, 45]. These methods test different sets of parameter values until they find the estimates that minimize the standardized distance to the observed data. FIML estimates the parameters in the likelihood directly from the available data in a way that does not require individuals to have complete data [35]. Parameter estimates obtained using the

FIML method are unbiased if data are MAR. This method does not use imputation. However, the FIML method relies on a family of parametric probability model assumptions, and correct model specification [32, 46]; when these assumptions are not tenable, the FIML method may result in less than optimal performance [47].

The development of an empirical method that explores the relationships that exist among items to address item non-response, without making any distributional assumption, would be beneficial to patient-oriented research. One potentially useful method that identifies the latent structure of observed variables and uses this information to address item non-response is non-negative matrix factorization (NNMF) [48, 49].

NNMF is an unsupervised machine-learning method that uses optimization techniques to find a low-dimensional data structure from the original data. It has been used for dimension reduction [50, 51] and feature extraction [52, 53] for high-dimensional data. NNMF has been applied in recommender systems to predict movie ratings from incomplete rating matrices [54–56]. These predictions can be recast as missing data imputation or matrix completion problems. Thus, NNMF is a potentially useful approach to address non-response at the item-level of PROMs data, which are ordinal and multivariate.

The NNMF, MI and FIML methods yield unbiased estimates under the MAR mechanism [45, 57]. When data are MNAR, results from these methods can be misleading. This is because parameters that describe the probability of non-response in the data are not estimated using these methods. One approach to address this problem is to use methods such as pattern mixture models, shared parameter models, and selection models designed for MNAR mechanism [34, 58]. However, these methods are not widespread in statistical software and are computationally intensive. Another approach is to ensure that the assumption of MAR mechanism is realistic in

the data [59]. This can be achieved by including variables that are potential correlates of missingness in the data [59]. These variables are called auxiliary or supplementary variables. Incorporating auxiliary variables into the missing data methods can help mitigate bias and improve statistical power [32, 60].

1.2 Purpose and Objectives

The research purpose was to develop and evaluate methods for addressing item non-response in PROMs. The objectives were to:

1. compare the performance of several ordinal item non-response methods when testing for differential item functioning.
2. evaluate the performance of several ordinal item non-response methods when estimating change in latent variable means over time.
3. estimate longitudinal change in PROMs scores in the presence of missing data using true-score model.
4. develop a new method to address ordinal item non-response and compare its performance with several ordinal item non-response methods.

1.3 Organization of Thesis

This thesis is comprised of four related manuscripts. Manuscript 1, "*A comparison of methods to address item non-response when testing for differential item functioning in multidimensional patient-reported outcome measures*", is presented in Chapter 2. This manuscript compares the performance of the NNMF method with FIML and listwise deletion when testing for differential item functioning in multidimensional PROMs using Monte Carlo simulation and real-world data. We found that the NNMF method demonstrated comparable performance to the FIML method.

Manuscript 2 is presented in Chapter 3; it is entitled "*Estimating longitudinal change in latent variable means: a comparison of non-negative matrix factorization and other item non-response methods*". This manuscript compares the performance of the NNMF method with the FIML, MI with conditional proportional odds model and listwise deletion when estimating longitudinal change in latent variable means. Simulated and real-world data were used to compare these methods. We found that the NNMF is relatively efficient when sample size is large and the percentage of non-response is high, but less optimal under other data-analytic conditions.

Chapter 4 motivates and presents the use of auxiliary variable in imputation models. Manuscript 3 is presented in Chapter 5, "*Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry"*. This manuscript investigates the use of auxiliary variables in imputation model and compared the precision and bias of FIML, MI with and without auxiliary variable, when estimating longitudinal change in PROM scores. We found that including auxiliary information in the imputation model increased the precision and reduced the bias of the estimated parameter.

Manuscript 4 is presented in Chapter 6, "*Enhanced weighted non-negative matrix factorization approach for addressing item non-response in patient-reported outcome measures*". This manuscript proposes an improved weighted NNMF, which uses information from an auxiliary variable to define strata and weights for missing or observed item responses before imputation, and compares the performance with FIML, NNMF and listwise deletion. Similar to other manuscripts, we used simulated and real-world data to compare these methods. The proposed showed comparable performance to the FIML method but outperformed the NNMF method for most of the simulation conditions considered.

Chapter 7 summarizes the research key findings, limitations, conclusions, and future directions.

1.4 References

1. Berzon, R., Hays, R. D., & Shumaker, S. A. (1993). International use, application and performance of health-related quality of life instruments. *Quality of Life Research*, *2*(6), 367–368.

2. Deshpande, P. R., Rajan, S., Sudeepthi, B. L., & Abdul Nazir, C. P. (2011). Patient-reported outcomes: a new era in clinical research. *Perspectives in Clinical Research*, *2*(4), 137–144.

3. Devji, T., Johnston, B. C., Patrick, D. L., Bhandari, M., Thabane, L., & Guyatt, G. H. (2017). Presentation approaches for enhancing interpretability of patient-reported outcomes (PROs) in meta-analysis: a protocol for a systematic survey of Cochrane reviews. *BMJ Open*, *7*(9), e017138.

4. Barclay, R., & Tate, R. B. (2014). Response shift recalibration and reprioritization in health-related quality of life was identified prospectively in older men with and without stroke. *Journal of Clinical Epidemiology*, *67*(5), 500–507.

5. Sprangers, M. A. G. (2002). Quality-of-life assessment in oncology. Achievements and challenges. *Acta Oncologica*, *41*(3), 229–237.

6. Estabrook, R., Cella, D., Zhao, F., Manola, J., DiPaola, R. S., Wagner, L. I., & Haas, N. B. (2018). Longitudinal and dynamic measurement invariance of the FACIT-fatigue scale: an application of the measurement model of derivatives to ECOG-ACRIN study E2805. *Quality of Life Research*, *27*(6), 1589–1597.

7. Jabrayilov, R., Emons, W. H. M., de Jong, K., & Sijtsma, K. (2017). Longitudinal measurement invariance of the Dutch outcome questionnaire-45 in a clinical sample. *Quality of Life Research*, *26*(6), 1473–1481.

8. Johnston, B. C., Patrick, D. L., Thorlund, K., Busse, J. W., da Costa, B. R., Schünemann, H. J., & Guyatt, G. H. (2013). Patient-reported outcomes in meta-analyses - part 2: methods for improving interpretability for decision-makers. *Health and Quality of Life Outcomes*, *11*(211), 1–9.

9. Guyatt, G. H., Feeny, D. H., & Patrick, D. L. (1993). Measuring health-related quality of life. *Annals of Internal Medicine*, *118*(8), 622–629.

10. Ware, J. E. Jr., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36) I. Conceptual framework and item selection. *Medical Care*, *30*(6), 473–483.

11. Testa, M. A., & Simonson, D. (1996). Assessment of quality-of-life outcomes. *New England Journal of Medicine*, *334*(13), 835–840.

12. Dunbar, M., Robertsson, O., Ryd, L. (2000). Translation and validation of the Oxford-12 item knee score for use in Sweden. *Acta Orthopaedica Scandinavica*, *71*(3), 268–274.

13. Dawson, J., Fitzpatrick, R., Carr, A., & Murray, D. (1996). Questionnaire on the perceptions of patients about total hip replacement. *Journal of Bone Joint Surgery*, *78-B*, 185–190.

14. Fayers, P. M., & Machin, D. (2007). *Quality of life: the assessment, analysis and interpretation of patient-reported outcomes* (Second.). Chichester: John Wiley & Sons, Ltd.

15. Bhandari, N. R., Kathe, N., Hayes, C., & Payakachat, N. (2018). Reliability and validity of SF-12v2 among adults with self-reported cancer. *Research in Social and Administrative Pharmacy*, *14*(11), 1080–1084.

16. Kathe, N., Hayes, C. J., Bhandari, N. R., & Payakachat, N. (2018). Assessment of reliability and validity of SF-12v2 among a diabetic population. *Value in Health*, *21*(4), 432–440.

17. Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543.

18. Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge/Taylor & Francis Group.

19. Liu, Y., Millsap, R. E., West, S. G., Tein, J. Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods*, *22*(3), 486–506.

20. Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data: an introduction. *Child Development Perspectives*, *4*(1), 5–9.

21. Wirth, R. J. (2008). *The effects of measurement non-invariance on parameter estimation in latent growth models*. University of North Carolina at Chapel Hill.

22. Meade, A. W., Lautenschlager, G. J., & Hecht, J. E. (2009). Establishing measurement equivalence and invariance in longitudinal data with item response theory. *International Journal of Testing*, *5*(3), 279–300.

23. Wu, X., Sawatzky, R., Hopman, W., Mayo, N., Sajobi, T. T., Liu, J., … Lix, L. M. (2017). Latent variable mixture models to test for differential item functioning: a population-based analysis. *Health and Quality of Life Outcomes*, *15*(1), 1–13.

24. Teresi, J. A. (2006). Different approaches to differential item functioning in health applications: advantages, disadvantages and some neglected topics. *Medical Care*, *44*(11 SUPPL. 3), 152–170.

25. Yadegari, I., Bohm, E., Ayilara, O. F., Zhang, L., Sawatzky, R., Sajobi, T. T., & Lix, L. M. (2019). Differential item functioning of the SF-12 in a population-based regional joint replacement registry. *Health and Quality of Life Outcomes*, *17*(1), 1–11.

26. Kwon, J. Y., & Sawatzky, R. (2017). Examining gender-related differential item functioning of the veterans rand 12-item health survey. *Quality of Life Research*, *26*(10), 2877–2883.

27. Lopez Rivas, G. E., Stark, S., & Chernyshenko, O. S. (2009). The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test. *Applied Psychological Measurement*, *33*(4), 251–265.

28. de Ayala, R. J. (2009). *The theory and practice of item response theory*. (D. Kenny, Ed.). New York: The Guilford Press.

29. Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373.

30. Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologist*. New Jersey: Lawrence Erlbaum Associates, Inc.

31. Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., … Stern, H. (2012). The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, *367*(14), 1355–1360.

32. Bell, M. L., & Fairclough, D. L. (2014). Practical and statistical issues in missing data for longitudinal patient-reported outcomes. *Statistical Methods in Medical Research*, *23*(5), 440–459.

33. Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *American Statistical Association*, *90*(431), 1112–1121.

34. Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

35. Enders, C. K. (2010). *Applied missing data analysis*. *The Guilford Press*. New York.

36. Schafer, J. L. (1997). *Analysis of incomplete multivariate data* (1st ed.). London, United Kingdom: Chapman & Hall Ltd.

37. van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*(12), 1049–1064.

38. Raghunathan, T. E., Lepkowski, J. M., & van Hoewyk, J. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27*(1), 85–95.

39. Donneau, A. F., Mauer, M., Molenberghs, G., & Albert, A. (2015). A simulation study comparing multiple imputation methods for incomplete longitudinal ordinal data. *Communications in Statistics: Simulation and Computation*, *44*(5), 1311–1338.

40. van Buuren, S. (2012). *Flexible imputation of missing data* (2nd ed.). Boca Raton, FL: CRC Press.

41. van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67.

42. van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, *16*(3), 219–242.

43. Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: a data analyst's perspective. *Multivariate Behavioral Research*, *33*(4), 545–571.

44. Nguyen, C. D., Carlin, J. B., & Lee, K. J. (2017). Model checking in multiple imputation: an overview and case study. *Emerging Themes in Epidemiology*, *14*(1), 1–12.

45. Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, *7*(2), 147–177.

46. Donneau, A. F., Mauer, M., Lambert, P., Molenberghs, G., & Albert, A. (2015). Simulation-based study comparing multiple imputation methods for non-monotone missing ordinal data in longitudinal settings. *Journal of Biopharmaceutical Statistics*, *25*(3), 570–601.

47. Baker, F. B., & Kim, S. H. (2004). *Item response theory: parameter estimation techniques*. (F. B. Baker & S.-H. Kim, Eds.) (2nd ed.). Boca Raton: CRC Press.

48. Aggarwal, C., & Parthasarathy, S. (2001). Mining massively incomplete data sets by conceptual reconstruction. In *ACM KDD* (pp. 227–232).

49. Agarwal, D., & Chen, B. C. (2009). Regression-based latent factor models. In *ACM KDD International Conference on Knowledge Discovery and Data Mining* (pp. 19–28).

50. Taslaman, L., & Nilsson, B. (2012). A framework for regularized non-negative matrix factorization, with application to the analysis of gene expression data. *PLoS ONE*, *7*(11), 1–7.

51. Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In T. K. Leen, T. G. Dietterich, & V. P. Tresp (Eds.), *Advances in Neural Information Processing Systems 13.* (pp. 556–562).

52. Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791.

53. Pauca, V. P., Piper, J., & Plemmons, R. J. (2006). Nonnegative matrix factorization for spectral data analysis. *Linear Algebra and Its Applications*, *416*(1), 29–47.

54. Zhang, S., Wang, W., Ford, J., & Makedon, F. (2006). Learning from incomplete ratings using non-negative matrix factorization. In *Proceedings of the Sixth SIAM International Conference on Data Mining* (pp. 549–553).

55. Mazumder, R., Hastie, T., & Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, *11*, 2287–2322.

56. Wold, H. (1975). Soft modelling by latent variables: the nonlinear iterative partial least squares (NIPALS) approach. *Journal of Applied Probability*, *12*(S1), 117–142.

57. Lin, X. E., & Boutros, P. C. (2020). Optimization and expansion of non-negative matrix factorization. *BMC Bioinformatics*, *21*(1), 1–10.

58. Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, *88*(421), 125–134.

59. Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351.

60. Eekhout, I., Enders, C. K., Twisk, J. W. R., de Boer, M. R., de Vet, H. C. W., & Heymans, M. W. (2015). Analyzing incomplete item scores in longitudinal data by including item score information as auxiliary variables. *Structural Equation Modeling*, *22*(4), 588–602.

CHAPTER 2: A COMPARISON OF METHODS TO ADDRESS ITEM NON-RESPONSE WHEN TESTING FOR DIFFERENTIAL ITEM FUNCTIONING IN MULTIDIMENSIONAL PATIENT-REPORTED OUTCOME MEASURES

2.1 Overview

In this research, the performance of several ordinal item non-response methods including non-negative matrix factorization, listwise deletion and full information maximum likelihood, when testing for differential item functioning in patient-reported outcome measures was investigated. Monte Carlo simulation and real-world data was used to compare the Type I error rate and statistical power of these methods. The non-negative matrix factorization method, which is an unsupervised machine-learning method that uses a set of optimization techniques to identify a low-dimensional data structure from the original data, demonstrated a comparable performance to the full information maximum likelihood method. This study is a unique addition to the literature on methods to address incomplete responses in patient-reported outcome measures because it focuses on ordinal items (i.e., Likert scale) and missing data. Patient-oriented research with a focus on the item level has not been extensively studied in patient-reported outcome measures. This is partly due to the increased computational time and large sample size required to model the relationship between latent variable and items. Investigating psychometric properties at the item level can provide an understanding of the impact of each item to the subscale or scale.

Publication Details

Ayilara, O.F., Sajobi, T.T., Barclay, R., Bohm, E., Jafari Jozani, M., & Lix, L.M. (2022). A comparison of methods to address item non-response when testing for differential item functioning in multidimensional patient-reported outcome measures. *Quality of Life Research*, 31, 2837-2848.

2.2 Abstract

Purpose: Item non-response (i.e., missing data) may mask the detection of differential item functioning (DIF) in patient-reported outcome measures (PROMs) or result in biased DIF estimates. Non-response can be challenging to address in ordinal data. We investigated an unsupervised machine-learning method for ordinal item-level imputation and compared it with commonly-used item non-response methods when testing for DIF.

Methods: Computer simulation and real-world data were used to assess several item non-response methods using the item response theory likelihood ratio test for DIF. The methods included: (a) listwise deletion (LD), (b) half-mean imputation (HMI), (c) full information maximum likelihood (FIML), and (d) non-negative matrix factorization (NNMF), which adopts a machine-learning approach to impute missing values. Control of Type I error rates were evaluated using a liberal robustness criterion for $\alpha = 0.05$ (i.e., 0.025 – 0.075). Statistical power was assessed with and without adoption of an item non-response method; differences >10% were considered substantial.

Results: Type I error rates for detecting DIF using LD, FIML and NNMF methods were controlled within the bounds of the robustness criterion for > 95% of simulation conditions, although the NNMF occasionally resulted in inflated rates. The HMI method always resulted in inflated error rates with 50% missing data. Differences in power to detect moderate DIF effects

for LD, FIML and NNMF methods were substantial with 50% missing data and otherwise insubstantial.

Conclusion: The NNMF method demonstrated comparable performance to commonly-used non-response methods. This computationally-efficient method represents a promising approach to address item-level non-response when testing for DIF.

2.3 Introduction

Patient-reported outcome measures (PROMs) are used to measure quality of life and other aspects of health and well-being [1–3]. Valid and reliable group comparisons of PROM scores rest on the assumption of equivalence of the item response patterns across groups within the population, after controlling for the underlying latent (i.e., unobserved) trait(s) of interest (e.g., quality of life). Differential item functioning (DIF) may affect the validity of group comparisons. DIF occurs when the probability of item response varies across groups defined by such characteristics as age or sex, after adjusting for one or more latent traits represented in the items.

Item non-response can mask the detection of DIF in PROMs, due to a loss of statistical power. Item non-response may be a particular problem in computerized adaptive testing (CAT) systems for PROMs. The design of a CAT system allows respondents to skip items or respond to items based on their preferences [4–6]. Previous studies about item non-response methods for PROMs data often assume the data follow a normal distribution [5]. This assumption is likely to be violated in ordinal data, particularly when the number of item response options is small.

Item non-response may occur because respondents are too sick to answer the questions, have difficulty understanding or interpreting the questions, are uncomfortable answering the questions, or miss scheduled clinic visits. The potential consequences of item-level non-response

include reduced statistical power and either over- or under-estimation of group differences [6, 7]. These effects could occur for tests of both uniform and non-uniform DIF [8].

No single optimal method has been recommended to address ordinal missing data [5]. The choice of methods is dependent, in part, on the missing data mechanism, which includes missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR) [6]. While listwise deletion (LD) [9] and half-mean imputation (HMI) [7] are common methods, they may result in biased tests of DIF, especially with a large amount of non-response [10]. Other methods to address ordinal item non-response include multiple imputation (MI) [11–14] and full information maximum likelihood (FIML) [6, 15]. For DIF analyses that adopt MI, the parameter estimates, standard errors and the model fit statistics such as the χ^2 goodness-of-fit statistic, root mean square error of approximation (RMSEA), comparative fit index (CFI), and standardized root mean squared residual (SRMR) are estimated for each imputed dataset. This can substantially increase computational time, especially with large numbers of imputations. Also, methods to pool model fit statistics from imputed datasets have not been explicitly defined [16, 17]. FIML rests on assumptions that may not be tenable in PROMs items [7, 18]; violation of these assumptions may result in less than optimal performance [19].

Non-negative matrix factorization (NNMF), an unsupervised machine-learning method that uses optimization techniques to find a low-dimensional structure from the original data, is a potentially useful method to address item non-response. Unsupervised machine-learning methods identify latent patterns in the data. In the context of missing data, observed information from individuals in these latent patterns are used to impute responses [20]. NNMF has been applied in recommender systems to predict movie ratings from incomplete rating matrices [21–23]. These predictions can be recast as missing data imputation or matrix completion problems. Thus,

NNMF is a potentially useful and novel approach to address non-response in item-level PROMs data, which are ordinal and multivariate (i.e., multiple correlated items).

Our purpose was to investigate NNMF as an ordinal item-level missing data imputation method. Our objectives were to (1) compare the performance of the NNMF method with the FIML, LD and HMI methods when testing for DIF using simulated data, and (2) evaluate the sensitivity of the NNMF method when compared with the FIML, LD and HMI methods when testing for DIF in real-world data.

2.4 Methods

2.4.1 Item Non-Response Methods

We begin with an overview of the investigated methods. The LD method removes all individuals with at least one item with a missing response. The HMI method fills in missing item responses with the mean of the available item responses if 50% or more of the item responses are complete and removes individuals with more than 50% missing responses on all the items [12, 24]. The FIML method is related to the maximum likelihood method, which chooses parameter values that assign the maximum possible probability or probability density to the observed data under a well-defined family of parametric probability models. The FIML method performs better than other item non-response methods, including random forest and logistic regression models appropriate for ordinal data [5]. However, it assumes that the latent trait follows a normal distribution, data are MAR, and correct model specification [25, 26]. The MI and FIML methods have similar performance [26–28], which is one reason we did not investigate the MI method.

For the NNMF method, suppose we have a $p \times n$ (p is the number of items, n is the number of respondents) response matrix $\boldsymbol{R}_{p\times n}$ with low rank k [$k \ll \min(n, p)$; $\ll$ means "much less

than"]. The NNMF method (Fig. 1) produces an item factor matrix $\boldsymbol{U}_{p\times k}$ and a respondent factor matrix $\boldsymbol{V}_{k\times n}$ such that $\boldsymbol{R} \approx \boldsymbol{UV}$, constraining $\boldsymbol{U}$ and $\boldsymbol{V}$ to be non-negative [29, 30]. The item factor matrix shows the affinity of each item for the latent trait(s), and the respondent factor matrix shows the affinity of each respondent for the latent trait(s). Decomposing the entire response matrix into smaller matrices allows the NNMF method to accurately capture dependencies amongst items [20]. Using this decomposition, one can reconstruct an approximation to the original data points in $\boldsymbol{R}$. However, when the missing data are MNAR, the resulting reconstruction may be biased [20]. Additional information about the NNMF method is provided in Online Resource 1. This method was implemented in RStudio version 3.5.2 using the *NNLM* package [20, 31].

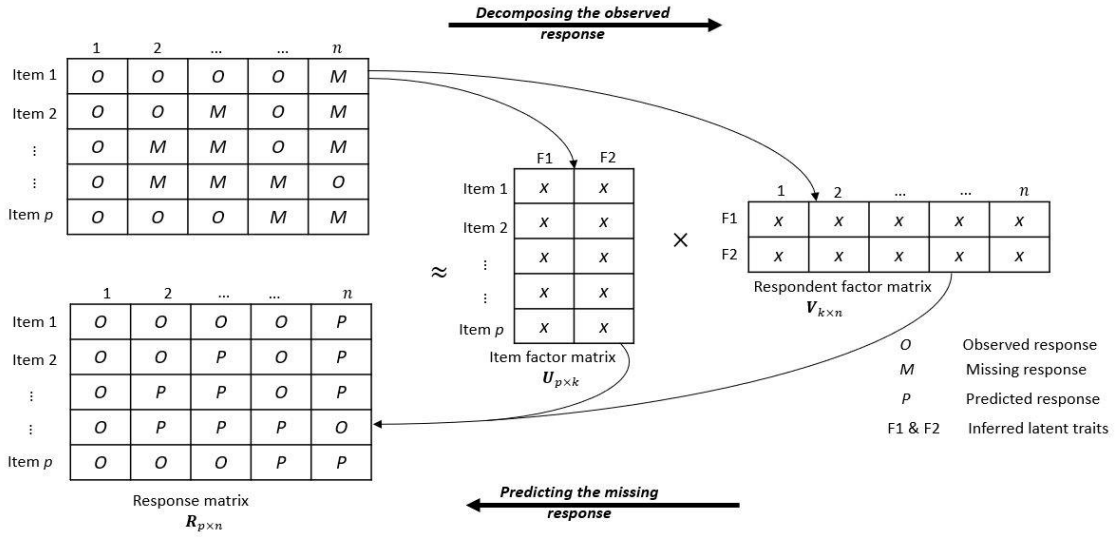


Figure 2-1. Visual illustration of the NNMF method

2.4.2 Simulation Study

Data generation

We simulated two-dimensional polytomous data for 30 items using the multidimensional graded response model (MGRM) [32, 33]; a description of the MGRM is provided in Online Resource

2. The first 15 items were associated with the first latent trait and the remaining 15 items were associated with the second latent trait. The slopes ($\boldsymbol{\alpha}$) of the items were sampled from a uniform distribution $U(1.1, 2.8)$ [33]. The thresholds ($\boldsymbol{\beta}$) also followed a uniform distribution $U(-2, 2)$ [34, 35]. The number of item response categories (c) was set to $c = 5$. The latent traits ($\boldsymbol{\theta}$) followed a multivariate normal distribution, $MVN(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the variance-covariance matrix.

Only the items that exhibited DIF had missing values, consistent with previous simulation studies [10, 36, 37]. The missingness on these items was achieved using the multivariate amputation method [38] that generates missing data by assigning a weight to each observation. This method has been shown to be more appropriate for generating missing data than a stepwise univariate amputation method (i.e., missingness on items are generated one item at a time) [39, 40]. Pre-specified amounts of responses were removed from the DIF items for each missing data mechanism using the multivariate *ampute* function in the *mice* package [41] in RStudio version 3.5.2.

Simulation conditions

The simulation conditions included sample size, percentage of missing data, magnitude of uniform and non-uniform DIF effects, and magnitude of the correlation between latent traits. Random samples of equal group sizes *n* = 250, 500 and 1000 were generated for each of the reference (R) and focal (F) groups and denoted as R250/F250, R500/F500, and R1000/F1000 giving total sample sizes that ranged from 500 to 2000 [9, 10, 42]. Random samples of unequal group sizes R1500/F500 were also investigated. Equal sizes were selected because they have been the focus of previous simulation studies [41]; unequal sizes were selected because they frequently occur in DIF analyses of real-world data [43–46]. Pre-specified amounts (0%, 10%,

25% and 50%) of cases were set to have missing responses on the items for both groups via MCAR, MAR, and MNAR mechanisms. The fully observed condition (i.e., 0% missing responses) was used to establish a baseline for assessing performance.

The percentage of items selected to exhibit DIF effects (i.e., DIF target items) was specified as 20% and 50%. The magnitude of the difference between the groups' thresholds and slopes were set to 0 (i.e., no DIF present), 0.2 (small), 0.4 (moderate) and 0.6 (large) [17]. The correlation between the two latent traits was set to ρ = 0.3 and 0.5. The diagonal elements of $\mathbf{\Sigma}$ were set to 1, while the off-diagonal elements were equal to ρ [36]. A total of 500 replications were conducted for each of 324 combinations of simulation conditions.

Simulation scenarios

First, we considered a scenario where neither uniform nor non-uniform DIF were present in the simulated data. No DIF effect was added to the threshold or slope parameters of the DIF target items. Next, we considered two scenarios where DIF was present. In scenarios where only uniform DIF was present, we added small to large DIF effects to the threshold parameters of the DIF target items in the focal group. In scenarios where non-uniform DIF was present, we added small to large DIF effects to the slope parameters of the DIF target items in the focal group.

Testing for DIF

We used the item response theory likelihood ratio test (IRT-LRT) to test for uniform and non-uniform DIF because of its popularity and availability in statistical software packages such as RStudio or MPlus. Other methods such as the multidimensional simultaneous item bias test (MULTISIB) [47] and the multidimensional multiple indicators multiple causes (MIMIC) interaction model can be used. The IRT-LRT and the MIMIC interaction model have shown to perform similarly [42]. The IRT-LRT compares the difference in the likelihood between two

nested MGRMs using a χ^2 statistic, with degrees of freedom equal to the difference in degrees of freedom between the two MGRMs [36, 42].

To test for DIF when it was not present, we fit a DIF model that assumed all items, except the target items were invariant between the focal and reference groups. Next, we fit a set of no-DIF models that constrained the threshold or slope parameters to be equal for all items in the two groups. A statistically significant LRT indicated the presence of DIF.

To test for uniform or non-uniform DIF when it was present, we fit a DIF model as specified above. Next, we fit no-DIF models that constrained the threshold or slope parameters to be equal for all items in the two groups. A statistically significant LRT when the DIF model was compared to: (1) a no-DIF model that constrained the threshold parameters for all items in the two groups indicated the presence of uniform DIF, and (2) a no-DIF model, where the slope parameters were constrained for all items in the two groups indicated the presence of non-uniform DIF. The MGRM-LRT was implemented in RStudio version 3.5.2. using the *mirt* package [48].

Performance measures

The performance of the item non-response methods was evaluated using the power rate (i.e., correctly identifying DIF) and Type I error rate (i.e., incorrectly identifying DIF). We report the average percentage of statistically significant LRT statistics across replications and items that were tested for DIF.

The robustness of Type I error rates was evaluated using Bradley's [49] liberal criterion (i.e., 0.025 – 0.075) for $\alpha = 0.05$. We compared the power of the MGRM-LRT with and without

adoption of a missing data method; differences in the power of the LRT ≤10% were considered insubstantial while differences >10% were considered substantial [50, 51].

2.4.3 Analyses of Real-World Data

Data source

Real-world data to demonstrate the implementation of the methods were from the population-based Winnipeg Regional Health Authority (WRHA) Joint Replacement Registry (JRR), from Manitoba, Canada. This registry contains both general-purpose and condition-specific PROMs [52] for patients having total or partial hip or knee arthroplasty surgeries. We focused on patients who had total hip arthroplasty between April 1, 2009, and March 31, 2015, the most recent data available at the time of the analysis.

Measures

The JRR captures the Short-Form 12 (SF-12), a general-purpose PROM consisting of twelve items that comprise physical component summary (PCS) and mental component summary (MCS) domains, each consisting of six items. The reliability and the validity of the SF-12 has been established in several populations [53–55].

Age group and sex were used to test for DIF consistent with previous studies [44, 45]. Age group was classified as ≤ 60 years and > 60 years.

Statistical analyses

We used stratified random sampling method to select patients from the entire study cohort, so that the size of the analytic cohort would be comparable to the largest total sample size in the simulation study. Samples of equal sizes were selected stratifying by sex and age group.

Descriptive statistics were used to characterize the study cohort. A three-step procedure was then used to test for uniform and non-uniform DIF using the MGRM with the LRT to: (1) assess SF-12 dimensionality, (2) select anchor items, and (3) assess DIF [8] (see Online Resource 2).

Model parameters were estimated using the maximum likelihood method. We used the Hochberg step-up approach to control the familywise Type I error rate to $\alpha = 0.05$ [56].

2.5 Results

2.5.1 Simulation Study Results

Fully observed item responses

Type I error and power rates for the LRTs of uniform and non-uniform DIF when the data were fully observed (i.e., no missing responses) are reported in Table 1. The results for the different percentages of items with DIF (20% versus 50%) were similar. The Type I error rates for all the simulation conditions were controlled within the bounds of Bradley's liberal criterion [49]. Group sizes and magnitude of DIF effects were positively associated with statistical power.

Table 2-1. Type I error rates (%) and power rates (%) to detect uniform and non-uniform DIF in fully observed data for the likelihood ratio test

% of DIF items	Sample size	ρ	Type I error	Power to detect uniform DIF			Power to detect non-uniform DIF		
				Magnitude of DIF effect					
				0.2	0.4	0.6	0.2	0.4	0.6
20	R1000/F1000	0.3	5.1	37.6	94.3	100.0	24.9	68.5	90.7
		0.5	5.7	35.5	93.2	100.0	35.6	74.9	93.7
	R500/F500	0.3	4.9	19.6	65.3	96.3	17.1	43.1	72.5
		0.5	5.2	19.5	66.6	95.9	20.0	50.3	74.7
	R250/F250	0.3	5.7	11.6	36.5	72.4	10.8	27.2	48.4
		0.5	5.2	13.1	37.6	71.3	13.4	29.8	49.4
50	R1000/F1000	0.3	5.2	35.4	92.7	100.0	28.6	71.0	93.3
		0.5	7.0	34.0	92.3	99.7	45.0	83.3	96.4
	R500/F500	0.3	4.1	17.9	64.0	94.5	17.0	42.2	69.3
		0.5	5.0	21.0	66.0	95.0	22.1	51.0	76.9
	R250/F250	0.3	4.6	10.3	35.5	72.6	12.0	26.9	44.9
		0.5	4.5	11.2	34.5	69.9	11.2	28.3	49.9

R: Reference group; F: Focal group; ρ: correlation between latent traits

Missing item responses

Table 2 reports the Type I error rates (i.e., percentages) for detecting DIF using the LRT when items contained missing responses and none of the 20% target items exhibited DIF effects. The results are stratified by latent trait correlation, percentage of missing responses and sample size. The Type I error rates for the LD and FIML methods were controlled within the bounds of the liberal criterion for all simulation conditions. The Type I error rates for the NNMF method were controlled within the bounds of the liberal criterion for 95.0% of simulation conditions. However, this method produced inflated error rates when sample size was small and the percentage of missing responses was 50%. The error rates of the HMI method often exceeded the upper bound of Bradley's liberal criterion, especially with 50% missing responses.

When sample sizes were unequal (i.e., R1500/F500), Type I error rates of all item non-response methods were slightly higher than when sample sizes were equal (R1000/F1000) when total sample size was held constant (see Online Resource 6).

Table 2-2. Type I error rates (%) for item non-response methods when testing for DIF using the likelihood ratio test when none of the 20% target items exhibited DIF effects

Mech.	Sample size	ρ	Percentage of missing responses											
			10%				25%				50%			
			LD	HMI	NNMF	FIML	LD	HMI	NNMF	FIML	LD	HMI	NNMF	FIML
MCAR	R1000/F1000	0.3	4.9	4.9	5.2	4.8	4.9	5.4	5.6	4.9	4.4	**8.8**	7.4	4.5
		0.5	6.1	6.3	6.3	6.1	5.6	**7.6**	6.9	5.5	5.8	**10.2**	7.5	5.8
	R500/F500	0.3	5.8	5.4	5.9	5.7	5.9	6.0	5.8	5.8	5.1	**9.0**	7.5	4.8
		0.5	4.8	5.8	5.5	4.8	5.4	6.1	5.8	5.1	5.4	**9.3**	7.4	5.5
	R250/F250	0.3	6.2	5.8	5.5	6.2	5.6	6.6	6.3	5.4	5.6	**11.2**	**8.5**	5.7
		0.5	5.4	5.5	5.4	5.4	5.2	6.5	5.5	5.2	5.6	**10.9**	**8.4**	5.5
MAR	R1000/F1000	0.3	5.2	4.8	4.7	5.1	4.7	5.4	5.3	4.9	4.4	**7.8**	6.6	4.4
		0.5	5.6	5.9	5.5	5.5	5.8	6.4	6.4	5.7	5.1	**9.7**	7.5	5.2
	R500/F500	0.3	5.9	6.3	6.2	5.7	4.8	6.9	5.6	4.8	6.6	**8.8**	**8.6**	6.1
		0.5	5.4	5.6	5.7	5.4	5.2	6.7	6.4	5.3	4.8	**11.1**	**7.6**	5.0
	R250/F250	0.3	5.6	5.6	5.4	5.9	5.7	6.9	6.7	5.6	5.6	**10.3**	**8.1**	5.4
		0.5	5.6	5.9	5.7	5.8	5.4	7.0	6.6	5.5	5.4	**10.1**	**8.1**	5.7
MNAR	R1000/F1000	0.3	4.7	5.0	4.6	4.7	5.0	6.2	5.4	4.9	4.8	**9.5**	6.9	5.0
		0.5	5.1	6.0	5.7	5.2	4.6	6.2	6.0	4.6	5.0	**9.5**	7.5	4.9
	R500/F500	0.3	5.8	6.3	5.8	5.7	6.0	6.6	6.1	5.7	4.0	**10.8**	**8.9**	4.2
		0.5	5.0	5.8	5.8	5.2	5.6	7.0	6.2	6.0	5.7	**8.9**	7.2	5.6
	R250/F250	0.3	5.6	6.0	5.6	5.5	5.6	6.4	5.9	5.6	5.6	**9.1**	7.5	5.6
		0.5	5.3	5.5	5.4	5.5	5.3	6.9	6.6	5.6	6.2	**10.6**	**7.6**	6.3

MCAR: Missing completely at random; MAR: Missing at random; MNAR: Missing not at random; LD: List-wise deletion; HMI: Half-mean imputation; NNMF: Non-negative matrix factorization; FIML: full information maximum likelihood; R: Reference group; F: Focal group; ρ: correlation between latent traits; Boldface font indicates error rates that fall outside the bounds of Bradley's [49] liberal criterion (i.e., 2.5 – 7.5%)

The power of the LRT for uniform DIF for latent trait correlations of 0.3 and 0.5 and sample size of R1000/F1000, stratified by magnitude of DIF and the missing data mechanism are presented in Figs. 2 and 3, respectively.

The power of the LRT decreased for all methods as the percentage of missing responses increased. The power to detect uniform DIF when 50% of the responses were missing ranged between 17.8% and 97.0% for the LD method; for the HMI method the corresponding values were 18.1% and 90.7%. For NNMF the corresponding values were 19.8% and 94.1%; and for FIML they were 17.2% and 97.1%. The power of the HMI method was more than 15 percentage points lower than the power of the LD, NNMF and FIML methods when the magnitude of the DIF effect was 0.4 and item responses were MNAR.

The power of the LRT to detect non-uniform DIF increased as the correlation between the latent traits increased from 0.3 to 0.5. When 50% of the responses were missing and sample size was R1000/F1000, power ranged between 14.9% and 74.1% for LD; for HMI the corresponding values were 21.1% and 78.2%. For NNMF the corresponding values were 19.3% and 74.9% and for FIML these values were 16.2% and 75.0%.

The differences in the power of the LRT to detect moderate (i.e., 0.4) and large (i.e., 0.6) uniform DIF effects when the missing mechanism was MNAR, and the percentage of non-response was 25% or higher for the HMI method were substantial. The power of the NNMF method was approximately 6 percentage points lower than the power of the LD, HMI and FIML methods when the magnitude of DIF was 0.6 and the latent trait correlation was 0.3. However, these differences decreased when the latent trait correlation increased to 0.5. The pattern of power results was similar for R500/F500 (see Online Resource 3) and R250/F250 (see Online Resource 4).

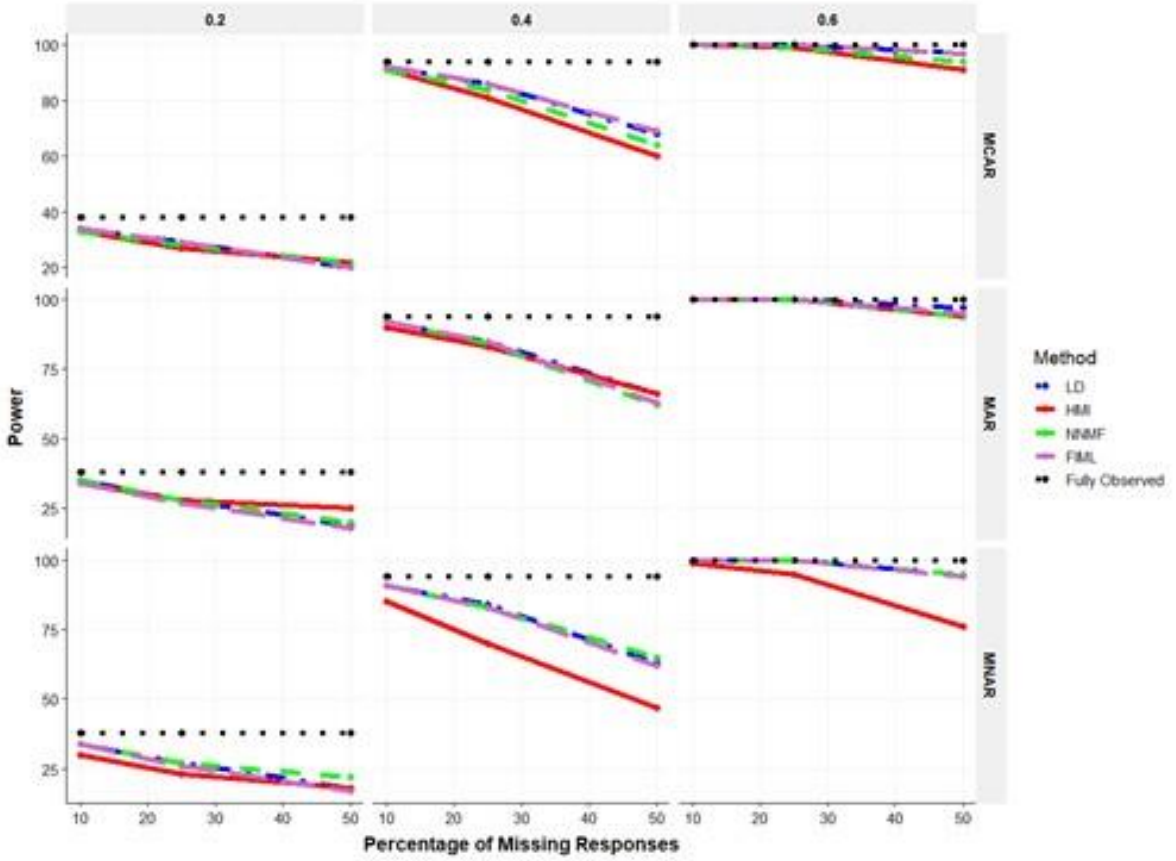


Figure 2-2. Power (%) of the LRT to detect uniform DIF for latent trait correlations of 0.3 and sample size of R1000/F1000 (R: reference group; F: focal group) stratified by the missing data mechanism and magnitude of DIF (20% of the items exhibited DIF). MCAR: missing completely at random; MAR: missing at random; and MNAR: missing not at random.

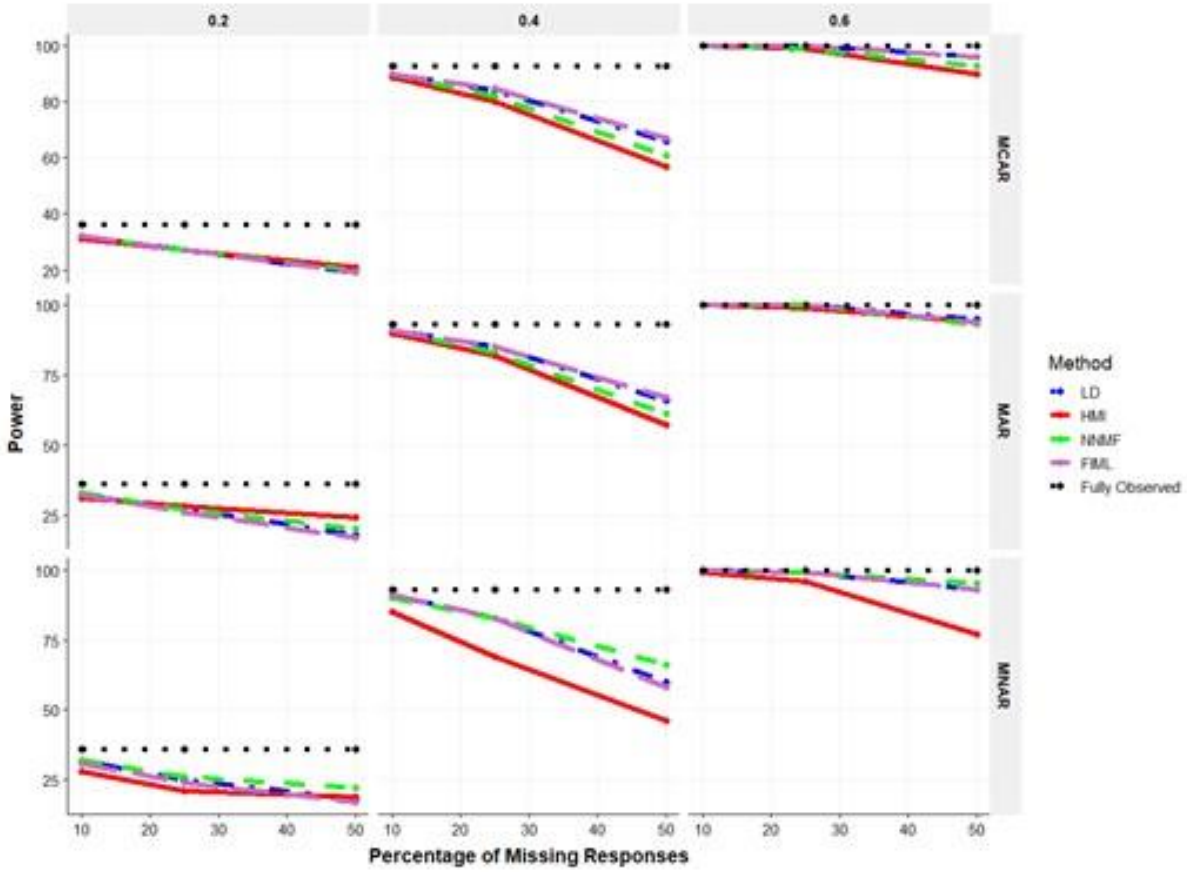


Figure 2-3. Power (%) of the LRT to detect uniform DIF for latent trait correlations of 0.5 and sample size of R1000/F1000 (R: reference group; F: focal group) stratified by the missing data mechanism and magnitude of DIF (20% of the items exhibited DIF). MCAR: missing completely at random; MAR: missing at random; and MNAR: missing not at random.

2.5.2 Real-World Data Analyses

The total JRR study cohort was comprised of 6469 patients. More than half (54.5%) were women and almost three-quarters (70.2%) were > 60 years. We randomly sampled 3000 patients stratified by sex and age group for the DIF analyses.

The distribution of responses for the PCS and MCS items are shown in Table 3. The average missing response rate per item (i.e., the sum of the percentage of missing responses for each item divided by the total number of items) was 12.7%. The highest percentages of missing responses

were 16.3% for PCS5 (Limited in work and other activities) and 17.4% for MCS2 (Did work or other activities less carefully than usual).

The first and second eigenvalues from the exploratory analysis had values of 6.00 and 1.73 respectively. The ratio of the eigenvalues was 3.47, which suggests a two-dimensional latent trait structure. The confirmatory analysis of a two-dimensional latent trait structure resulted in reasonable model fit (RMSEA = 0.059, 90% CI = 0.052–0.067, CFI = 0.98, SRMR = 0.10). Figure 4 shows the model parameter estimates. The correlation between the PCS and MCS latent variables was 0.58. The unstandardized estimated slopes for the PCS and MCS items with complete responses ranged from 0.52 to 5.35, and 1.10 to 5.67, respectively.

On the PCS, “accomplished less than you would like” and on the MCS “did work or other activities less carefully than usual”, had the highest estimated slopes amongst all items that were non-significant using the all-other anchor approach (AOAA) [53, 54]. These items were selected as anchor items.

Table 4 shows the difference in the LRT statistics for unconstrained and constrained DIF models for the PCS and MCS items for sex for each item non-response method. One PCS item exhibited uniform DIF for all methods. There was no evidence of non-uniform DIF for males and females for the six PCS items for all item non-response methods. Two MCS items exhibited uniform DIF by sex for all methods. One MCS item was identified as having non-uniform DIF for the HMI, FIML and NNMF methods, but there was no evidence of non-uniform DIF for the LD method. The results for tests of uniform and non-uniform DIF by age group are provided in Online Resource 4.

Table 2-3. Distribution of item responses (%) on the SF-12 physical component summary (PCS) and mental component summary (MCS) for patients having total hip arthroplasty (N = 3000)

Item	**Response options**					
Physical Component Summary (PCS)	**Excellent**	**Very good**	**Good**	**Fair**	**Poor**	**Missing**
PCS1: General health	52 (1.8)	282 (9.4)	1179 (39.3)	886 (29.5)	202 (6.7)	399 (13.3)
	Limited a lot	**Limited a little**	**Not limited at all**	**Missing**		
PCS2: Limits in moderate activity	1706 (56.9)	738 (24.6)	141 (4.7)	415 (13.8)		
PCS3: Climbing several flights	1845 (61.5)	625 (20.8)	110 (3.7)	420 (14.0)		
	All of the time	**Most of the time**	**Some of the time**	**A little of the time**	**None of the time**	**Missing**
PCS4: Accomplished less	890 (29.7)	926 (30.9)	496 (16.5)	163 (5.4)	65 (2.2)	460 (15.3)
PCS5: Limited in work and other activities	935 (31.2)	930 (31.0)	455 (15.2)	139 (4.6)	52 (1.7)	489 (16.3)
PCS6: Pain interference with normal work	24 (0.8)	173 (5.8)	521 (17.4)	1151 (38.4)	670 (22.3)	461 (15.4)
Mental Component Summary (MCS)						
MCS1: Accomplished less	228 (7.6)	416 (13.9)	612 (20.4)	521 (17.4)	750 (25.0)	473 (15.8)
MCS2: Less careful than usual	205 (6.8)	400 (13.3)	532 (17.7)	512 (17.1)	828 (27.6)	532 (17.4)
MCS3: Felt calm and peaceful	136 (4.5)	399 (13.3)	728 (24.3)	1127 (37.6)	170 (5.7)	440 (14.7)
MCS4: Have a lot of energy	361 (12.0)	695 (23.2)	911 (30.4)	511 (17.0)	72 (2.4)	450 (15.0)
MCS5: Felt downhearted and depressed	49 (1.6)	160 (5.3)	697 (23.2)	851 (28.4)	791 (26.4)	452 (15.1)
MCS6: Emotional problems	216 (7.2)	475 (15.8)	746 (24.9)	527 (17.6)	617 (20.6)	419 (14.0)

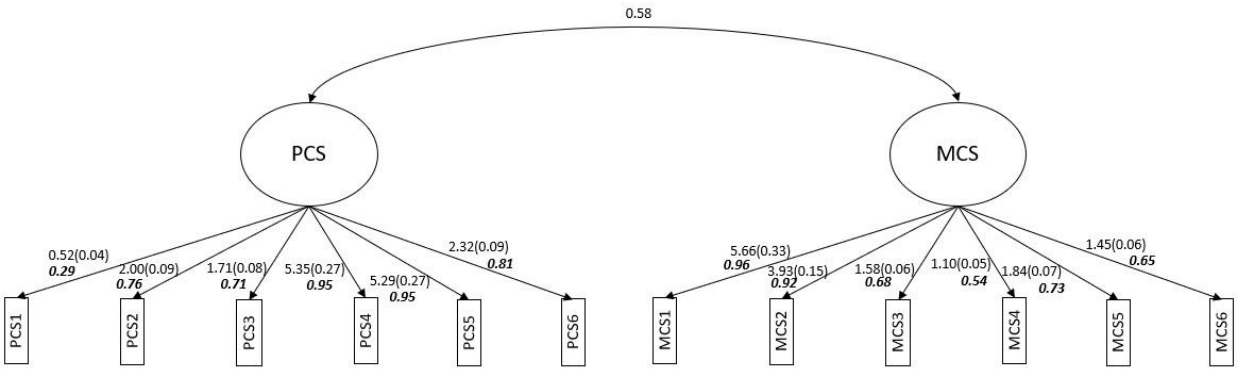


Figure 2-4. A two-dimensional graded response model for the physical component summary (PCS) and mental component summary (MCS) of the SF-12 in the real-world data. The numbers shown represent the unstandardized slopes with the associated standard errors (in parentheses), and standardized slopes in boldface font. Variances of the PCS and MCS were fixed to 1.

Table 2-4. Difference in the log-likelihood statistics of an unconstrained and constrained DIF model by sex (degrees of freedom) for the SF-12 physical component summary (PCS) and mental component summary (MCS) items in the real-world data for the item non-response methods

Item	**Uniform DIF**				**Non-uniform DIF**			
Physical Component Summary (PCS)	**LD**	**HMI**	**NNMF**	**FIML**	**LD**	**HMI**	**NNMF**	**FIML**
PCS1: General health	1.25 (4)	0.65 (4)	3.92 (4)	0.91 (4)	5.19 (1)	5.70 (1)	0.77 (1)	5.06 (1)
PCS2: Limits in moderate activity	3.35 (2)	3.37 (2)	1.34 (2)	3.77 (2)	0.55 (1)	0.14 (1)	0.20 (1)	0.07 (1)
PCS3: Climbing several flights	**13.32 (2)**	**13.24 (2)**	**11.49 (2)**	**11.89 (2)**	4.19 (1)	4.16 (1)	5.19 (1)	4.16 (1)
PCS4: Accomplished less	†	†	†	†	†	†	†	†
PCS5: Limited in work and other activities	6.96 (4)	2.87 (4)	3.80 (4)	3.63 (4)	5.07 (1)	1.55 (1)	1.75 (1)	2.11 (1)
PCS6: Pain interference with normal work	10.43 (4)	9.23 (4)	4.77 (4)	9.42 (4)	2.12 (1)	3.05 (1)	1.98 (1)	2.59 (1)
Mental Component Summary (MCS)								
MCS1: Accomplished less	12.57 (4)	13.19 (4)	13.67 (4)	13.87 (4)	4.44 (1)	2.53 (1)	1.29 (1)	3.81 (1)
MCS2: Less careful than usual	†	†	†	†	†	†	†	†
MCS3: Felt calm and peaceful	**19.38 (4)**	**22.62 (4)**	**27.88 (4)**	**20.32 (4)**	6.56 (1)	**10.38 (1)**	**10.34 (1)**	**8.61 (1)**
MCS4: Have a lot of energy	**57.72 (4)**	**54.10 (4)**	**48.91 (4)**	**56.98 (4)**	0.16 (1)	0.33 (1)	0.06 (1)	0.05 (1)
MCS5: Felt downhearted and depressed	8.48 (4)	8.89 (4)	9.14 (4)	8.18 (4)	0.05 (1)	0.39 (1)	0.52 (1)	0.20 (1)
MCS6: Emotional problems	12.56 (4)	11.62 (4)	12.90 (4)	12.05 (4)	0.31 (1)	0.00 (1)	0.01 (1)	0.10 (1)

†= anchor item; Boldface font indicates a statistically significant Hochberg adjusted p-value; LD: List-wise deletion; HMI: Half-mean imputation; NNMF: Non-negative matrix factorization; FIML: full information maximum likelihood

2.6 Discussion

This study investigated the NNMF method for addressing item non-response and compared it to the LD, HMI and FIML methods when testing for DIF in multidimensional PROMs via simulated and real-world data. We investigated the NNMF method because it has the advantage of being computationally efficient and makes no distributional assumptions about the item responses before imputation.

The results indicate that the NNMF method performed similar to the LD and FIML methods when sample size was large. However, when sample size was small and the percentage of missing responses was high (i.e., 50%), the NNMF method could result in inflated error rates. At the same time, the Type I error rates for the NNMF, LD and FIML methods were controlled within the bounds of Bradley's liberal criterion [49] for the vast majority of the investigated simulation conditions. The findings for the LD and FIML methods are consistent with previous studies [10, 37].

Type I error rates were slightly higher when sample sizes were unequal compared to equal. However, the NNMF, LD and FIML methods still produced error rates within the bounds of Bradley's liberal criterion [49]. The power to detect uniform and non-uniform DIF for the unequal sample size condition was slightly lower than when sample sizes were equal. These findings are consistent with a previous study on DIF detection [42].

The difference in power to detect moderate DIF effects for the LD, FIML and NNMF methods were substantial with 50% missing responses and insubstantial for less than 50% missing responses. As expected, power to detect uniform and non-uniform DIF increased as sample size and magnitude of DIF effects increased but decreased as the percentage of item non-response increased. This finding is also consistent with previous studies [10, 55, 56]. Power to detect non-

uniform DIF increased as the latent trait correlation increased; this trend was not observed for uniform DIF. Overall, the power to detect uniform DIF was greater than the power to detect non-uniform DIF.

In the analysis of the JRR data, the PCS item corresponding to climbing several flights showed evidence of uniform DIF for all item non-response methods. The MCS items corresponding to felt calm and peaceful, and having lots of energy showed evidence of uniform DIF across all methods. All methods were equally sensitive to detect DIF.

The NNMF method can be advantageous over other methods when producing summary information across items. The NNMF method is straightforward to implement because it uses the observed responses from all respondents to impute item-level missing data. In contrast, the LD method deletes respondents with at least one missing response, while the HMI method deletes respondents with more than 50% missing responses on all the items. Finally, the FIML method cannot be used in this case because item responses are not directly imputed.

There are, however, some challenges associated with using the NNMF method to address item non-response. First, this method performs poorly when the correlations amongst items are weak. The NNMF method relies on the associations amongst items to predict missing responses. Weak correlations imply the associations amongst items are not sufficiently strong to accurately predict missing responses [21, 22, 29]. Second, the low rank of the item response matrix needs to be specified. We specified a low rank of two because the SF-12 is often characterized as having two component domains. A data-driven method could be used to select the low rank when there is no prior information about the number of domains [20].

2.6.1 Strengths and Limitations

This study has a number of strengths. First, we used a combination of simulated and real-world data to investigate several item non-response methods for DIF detection in multidimensional PROMs items. The NNMF method was investigated because it uses the relationship amongst item responses and underlying latent traits to impute missing responses without making any distributional assumptions.

There are, however, some limitations to this study. First, like other simulation studies, the simulation conditions we investigated were not exhaustive. We limited the number of item response categories, number of items [17], and total sample size; larger numbers for these conditions would substantially increase computational time. We assumed all items were invariant except the target items, which implies a large number of anchor items relative to items that exhibited DIF. This may not be the case in real-world data. Also, we only considered group differences in item thresholds (uniform DIF) or slopes (non-uniform DIF) in the simulation study. Differences might occur in both item thresholds and slopes, although this is unlikely to affect the performance of the item non-response methods. In addition, the methods we compared to the NNMF method are best suited for MCAR and MAR cases. Therefore, a future study should compare the performance of the NNMF method with methods designed to address MNAR missing data.

2.6.2 Conclusion

In summary, we compared the performance of different item non-response methods when testing for uniform and non-uniform DIF in multidimensional PROMs data. While the NNMF method compared favorably to the LD and FIML methods, it resulted in inflated Type I error rates when

sample size decreased. Statistical power rates for all but the HMI method were reasonably close to those from fully observed data.

At the same time, we found no single method performed best in terms of statistical power. Consistent with previous studies, we recommend that researchers consider comparing two or more methods in sensitivity analyses [7, 57]. Previous studies that examined missing data methods and parameter recovery, but not DIF, have shown that the LD method resulted in biased parameter estimates and inflated standard errors [58–60]; therefore we do not recommend its use The FIML and NNMF methods are, accordingly, two reasonable choices to compare in sensitivity analyses.

Declarations

Competing interests. None declared

Availability of data and material. Study data for the real-world analyses were secondary data. These data were provided under specific data sharing agreements only for approved use for this project. The original source data are not owned by the researchers and as such cannot be provided to a public repository. Where necessary and with appropriate approvals, the original source data for this project may be reviewed with the consent of the data providers and approval by the required privacy and ethical review bodies.

Authors' contributions. All authors conceived the study and contributed to the design of the simulation study and analysis plan for the numeric example. OFA, MJJ and LML conducted the analysis and prepared the draft manuscript. All authors reviewed and approved the final version of the manuscript.

Ethics approval. This study received ethical approval from the University of Manitoba Health Research Ethics Board.

Consent to participate. Informed written consent was obtained from all participants whose information was used in the analyses of data from the Winnipeg Regional Health Authority Joint Replacement Registry.

License to publish. The corresponding author grants to the Springer Nature Switzerland AG the perpetual, exclusive, world-wide, assignable, sublicensable and unlimited right to: publish, reproduce, copy, distribute, communicate, display publicly, sell, rent and/or otherwise make available this article, including supplementary materials.

2.7 References

1. Johnston, B. C., Patrick, D. L., Thorlund, K., Busse, J. W., da Costa, B. R., Schünemann, H. J., & Guyatt, G. H. (2013). Patient-reported outcomes in meta-analyses --part 2: methods for improving interpretability for decision-makers. *Health and Quality of Life Outcomes*, *11*(211), 1–9.

2. Guyatt, G. H., Feeny, D. H., & Patrick, D. L. (1993). Measuring health-related quality of life. *Annals of Internal Medicine*, *118*(8), 622–629.

3. Berzon, R., Hays, R. D., & Shumaker, S. A. (1993). International use, application and performance of health-related quality of life instruments. *Quality of Life Research*, *2*(6), 367–368.

4. Bulut, O., & Kim, D. (2021). The use of data imputation when investigating dimensionality in sparse data from computerized adaptive tests. *Journal of Applied Testing Technology*, *22*(2).

5. Jia, F., & Wu, W. (2019). Evaluating methods for handling missing ordinal data in structural equation modeling. *Behavior Research Methods*, *51*(5), 2337–2355.

6. Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

7. Bell, M. L., & Fairclough, D. L. (2014). Practical and statistical issues in missing data for longitudinal patient-reported outcomes. *Statistical Methods in Medical Research*, *23*(5), 440–459.

8. Teresi, J. A., & Fleishman, J. A. (2007). Differential item functioning and health assessment. *Quality of Life Research*, *16*(SUPPL. 1), 33–42.

9. Banks, K. (2015). An introduction to missing data in the context of differential item functioning. *Practical Assessment, Research and Evaluation*, *20*(12), 1–10.

10. Finch, H. (2011). The use of multiple imputation for missing data in uniform DIF analysis: power and type I error rates. *Applied Measurement in Education*, *24*(4), 281–301.

11. Donneau, A. F., Mauer, M., Molenberghs, G., & Albert, A. (2015). A simulation study comparing multiple imputation methods for incomplete longitudinal ordinal data. *Communications in Statistics: Simulation and Computation*, *44*(5), 1311–1338.

12. Eekhout, I., De Vet, H. C. W., Twisk, J. W. R., Brand, J. P. L., De Boer, M. R., & Heymans, M. W. (2014). Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. *Journal of Clinical Epidemiology*, *67*(3), 335–342.

13. Kombo, A. Y., Mwambi, H., & Molenberghs, G. (2017). Multiple imputation for ordinal longitudinal data with monotone missing data patterns. *Journal of Applied Statistics*, *44*(2), 270–287.

14. Raghunathan, T. E., Lepkowski, J. M., & Van Hoewyk, J. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27*(1), 85–95.

15. Enders, C. K. (2010). *Applied missing data analysis*. *The Guilford Press*. New York.

16. Liu, Y., Millsap, R. E., West, S. G., Tein, J. Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods*, *22*(3), 486–506.

17. Chen, P. Y., Wu, W., Garnier-Villarreal, M., Kite, B. A., & Jia, F. (2020). Testing measurement invariance with ordinal missing data: a comparison of estimators and missing data techniques. *Multivariate Behavioral Research*, *55*(1), 87–101.

18. Donneau, A. F., Mauer, M., Lambert, P., Molenberghs, G., & Albert, A. (2015). Simulation-based study comparing multiple imputation methods for non-monotone missing ordinal data in longitudinal settings. *Journal of Biopharmaceutical Statistics*, *25*(3), 570–601.

19. Baker, F. B., & Kim, S. H. (2004). *Item response theory: parameter estimation techniques*. (F. B. Baker & S.-H. Kim, Eds.) (2nd ed.). Boca Raton: CRC Press.

20. Lin, X. E., & Boutros, P. C. (2020). Optimization and expansion of non-negative matrix factorization. *BMC Bioinformatics*, *21*(1), 1–10.

21. Zhang, S., Wang, W., Ford, J., & Makedon, F. (2006). Learning from incomplete ratings using non-negative matrix factorization. In *Proceedings of the Sixth SIAM International Conference on Data Mining* (pp. 549–553).

22. Mazumder, R., Hastie, T., & Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, *11*, 2287–2322.

23. Wold, H. (1975). Soft modelling by latent variables: the nonlinear iterative partial least squares (NIPALS) approach. *Journal of Applied Probability*, *12*(S1), 117–142.

24. Fairclough, A. D. L., & Cella, D. F. (1996). Functional assessment of cancer therapy (FACT-G): Non-response to individual questions. *Quality of Life Research*, *5*(3), 321–329.

25. Enders, C. K. (2004). The impact of missing data on sample reliability estimates: implications for reliability reporting practices. *Educational and Psychological Measurement*, *64*(3), 419–436.

26. Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351.

27. Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, *7*(2), 147–177.

28. Ayilara, O. F., Zhang, L., Sajobi, T. T., Sawatzky, R., Bohm, E., & Lix, L. M. (2019). Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. *Health and Quality of Life Outcomes*, *17*(1), 106.

29. Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791.

30. Pauca, V. P., Piper, J., & Plemmons, R. J. (2006). Nonnegative matrix factorization for spectral data analysis. *Linear Algebra and Its Applications*, *416*(1), 29–47.

31. Lin, X. E., & Boutros, P. (2019). NNLM: a package for fast and versatile nonnegative matrix factorization. Retrieved from https://github.com/linxihui/NNLM

32. Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: limited versus full information methods. *Psychological Methods*, *14*(3), 275–299.

33. Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in Psychology*, *7*(February).

34. Olsbjerg, M., & Christensen, K. B. (2015). Modeling local dependence in longitudinal IRT models. *Behavior Research Methods*, *47*(4), 1413–1424.

35. De Ayala, R. J. (1994). The influence of multidimensionality on the graded response model. *Applied Psychological Measurement*, *18*(2), 155–170.

36. Bulut, O., & Sunbul, Ö. (2017). Monte Carlo simulation studies in item response theory with the R programming language. *Journal of Measurement and Evaluation in Education and Psychology*, *8*(3), 266–287.

37. Finch, H. W. (2011). The impact of missing data on the detection of nonuniform differential item functioning. *Educational and Psychological Measurement*, *71*(4), 663–683.

38. Schouten, R. M., Lugtig, P., & Vink, G. (2018). Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation*, *88*(15), 2909–2930.

39. Nassiri, V., Molenberghs, G., Verbeke, G., & Barbosa-Breda, J. (2020). Iterative multiple imputation: a framework to determine the number of imputed datasets. *American Statistician*, *74*(2), 125–136.

40. Goretzko, D. (2021). Factor retention in exploratory factor analysis with missing data. *Educational and Psychological Measurement*, 1–21.

41. van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67.

42. Bulut, O., & Suh, Y. (2017). Detecting multidimensional differential item functioning with the multiple indicators multiple causes model, the item response theory likelihood ratio test, and logistic regression. *Frontiers in Education*, *2*(October), 1–14.

43. Bourion-Bédès, S., Schwan, R., Laprevote, V., Bédès, A., Bonnet, J. L., & Baumann, C. (2015). Differential item functioning (DIF) of SF-12 and Q-LES-Q-SF items among French substance users. *Health and Quality of Life Outcomes*, *13*(1).

44. Yadegari, I., Bohm, E., Ayilara, O. F., Zhang, L., Sawatzky, R., Sajobi, T. T., & Lix, L. M. (2019). Differential item functioning of the SF-12 in a population-based regional joint replacement registry. *Health and Quality of Life Outcomes*, *17*(1), 1–11. https://doi.org/10.1186/s12955-019-1166-1

45. Lix, L. M., Wu, X., Hopman, W., Mayo, N., Sajobi, T. T., Liu, J., … Sawatzky, R. (2016). Differential item functioning in the SF-36 physical functioning and mental health sub scales: a population-based investigation in the Canadian multicentre osteoporosis study. *PLoS ONE*, *11*(3), 1–13.

46. Kwon, J. Y., & Sawatzky, R. (2017). Examining gender-related differential item functioning of the veterans rand 12-item health survey. *Quality of Life Research*, *26*(10), 2877–2883.

47. Stout, W., Li, H. H., Nandakumar, R., & Bolt, D. (1997). MULTISIB: a procedure to investigate DIF when a test is intentionally two-dimensional. *Applied Psychological Measurement*, *21*(3), 195–213.

48. Chalmers, R. P. (2012). mirt: a multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29.

49. Bradley, J. V. (1978). Robustness. *British Journal of Mathematical & Statistical Psychology*, *31*(2), 144–152.

50. Kaplan, D. (1989). A study of the sampling variability and z-values of parameter estimates from misspecified structural equation models. *Multivariate Behavioral Research*, *24*(1), 41–57.

51. Curran, P., & West, S. G. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*(1), 16–29.

52. Zhang, L., Lix, L. M., Ayilara, O., Sawatzky, R., & Bohm, E. R. (2018). The effect of multimorbidity on changes in health-related quality of life following hip and knee arthroplasty. *Bone and Joint Journal*, *100B*(9), 1168–1174.

53. Salyers, M., Bosworth, H., Swanson, J., Lamb-Pagone, J., & Osher, F. (2000). Reliability and validity of the SF-12 health survey among people with severe mental illness. *Medical Care*, *38*, 1141–1150.

54. Cernin, P., Cresci, K., Jankowski, T., & Lichtenberg, P. (2010). Reliability and validity testing of the short-form health survey in a sample of community-dwelling African American older adults. *Journal of Nursing Measurement*, *18*, 49–59.

55. Cheak-Zamora, N., Wyrwich, K., & McBride, T. (2009). Reliability and validity of the SF-12v2 in the medical expenditure panel survey. *Quality of Life Research*, *18*, 727–735.

56. Hochberg Yosef. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, *75*(4), 800–802.

57. Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology*, *97*(5), 1016–1031.

58. Kopf, J., Zeileis, A., & Strobl, C. (2015). *Anchor selection strategies for DIF analysis: review, assessment, and new approaches*. *Educational and Psychological Measurement* (Vol. 75).

59. Sedivy, S. K., Zhang, B., & Traxel, N. M. (2006). Detection of differential item functioning with polytomous items in the presence of missing data. In *Annual meeting of the National Council on Measurement in Education*.

60. Finch, H. W. (2011). The impact of missing data on the detection of nonuniform differential item functioning. *Educational and Psychological Measurement*, *71*(4), 663–683.

61. Rombach, I., Rivero-Arias, O., Gray, A. M., Jenkinson, C., & Burke, Ó. (2016). The current practice of handling and reporting missing outcome data in eight widely used PROMs in RCT publications: a review of the current literature. *Quality of Life Research*, *25*(7), 1613–1623.

62. Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, *45*(3), 225–245.

63. Finch, W. H. (2010). Imputation methods for missing categorical questionnaire data: a comparison of approaches. *Journal of Data Science*, *8*(3), 361–378.

64. Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In T. K. Leen, T. G. Dietterich, & V. P. Tresp (Eds.), *Advances in Neural Information Processing Systems 13*. (pp. 556–562).

65. Slocum-Gori, S. L., & Zumbo, B. D. (2011). Assessing the unidimensionality of psychological scales: using multiple criteria from factor analysis. *Social Indicators Research*, *102*(3), 443–461.

66. Reise, S. P., Cook, K. F., & Moore, T. M. (2014). Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In S. P. Reise & D. A. Revicki. (Eds.), *Handbook of item response theory modeling: applications to typical performance assessment* (pp. 13–37). New York: Routledge.

67. Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21*(2), 230–258.

68. Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55.

69. Lopez Rivas, G. E., Stark, S., & Chernyshenko, O. S. (2009). The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test. *Applied Psychological Measurement, 33*(4), 251–265.

Supplementary Materials

A comparison of methods to address item non-response when testing for differential item functioning in multidimensional patient-reported outcome measures

Online Resource 1: Description of the non-negative matrix factorization (NNMF) method

Suppose we have a $p \times n$ (where p is the number of items and n is the number of respondents) response matrix $\boldsymbol{R}_{p\times n}$ with low rank k [$k \ll \min(n, p)$; $\ll$ means "much less than"]. The NNMF method produces an item factor matrix $\boldsymbol{U}_{p\times k}$ and a respondent factor matrix $\boldsymbol{V}_{k\times n}$ such that $\boldsymbol{R} \approx \boldsymbol{UV}$, constraining $\boldsymbol{U}$ and $\boldsymbol{V}$ to be non-negative. The NNMF method minimizes:

$$P(\boldsymbol{U}, \boldsymbol{V}) = \frac{1}{2} \|\boldsymbol{R} - \boldsymbol{UV}\|^2, \quad (1)$$

subject to $\boldsymbol{U} \geq 0, \boldsymbol{V} \geq 0$, where $\|\,.\,\|^2$ is the squared Frobenius norm of the matrix [64].

We used multiplicative update rules, an iterative method to update factor matrices $\boldsymbol{U}$ and $\boldsymbol{V}$, because of their ease of implementation and speed of convergence. The multiplicative update rules for entries u_{ik} and v_{kj} of the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ respectively, with the penalty terms as specified in Eq.(1) are:

$$u_{ik} \Leftarrow \frac{(\boldsymbol{RV}^{\mathrm{T}})_{ik} u_{ik}}{(\boldsymbol{UVV}^{\mathrm{T}})_{ik}}, \quad (2)$$

$$v_{kj} \Leftarrow \frac{(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{R})_{kj} v_{kj}}{(\boldsymbol{U}^{T}\boldsymbol{U}\,\boldsymbol{V})_{kj}}, \quad (3)$$

Now, consider the situation where some entries of $\boldsymbol{R}$ are missing. The initial formulation of the multiplicative update rules will be insufficient unless additional constraints are imposed on the rows and columns of $\boldsymbol{U}$ and $\boldsymbol{V}$. For an incomplete ordinal item response matrix with low rank k, let $\omega \subset \{1, \ldots, p\} \times \{1, \ldots, n\}$ denote the indices of observed entries, $I_j = \{i: r_{ij}$ is observed$\}$ and $\widetilde{I_j} = \{i: r_{ij}$ is missing$\}$. We observe that information in $\boldsymbol{R}$ is redundant to achieve $\boldsymbol{R} \approx \boldsymbol{UV}$ since $\boldsymbol{R}$ is assumed to have a low rank k. Therefore, factorization of the incomplete ordinal item

response matrix can be achieved using the observed response. The optimization problem in Eq.(1) can then be reformulated as,

$$\underset{U,V}{\text{minimize}} \ \frac{1}{2}\sum_{(i,j)\in\omega}(\boldsymbol{R}_{ij}-(\boldsymbol{UV})_{ij})^2, \tag{4}$$

subject to $\boldsymbol{U}\geq 0, \boldsymbol{V}\geq 0$. When applying the multiplicative update rules on the j^{th} column of $\boldsymbol{V}$ in Eq.(4), all $\widetilde{I_j}$ rows of $\boldsymbol{U}$ are removed. The resulting multiplicative update equation with squared Frobenius norm becomes,

$$v_{kj} \Leftarrow \frac{\left(\boldsymbol{U}_{I_{j.}}^{\mathrm{T}}\boldsymbol{R}_{I_{j.}}\right)_{kj} v_{kj}}{\left(\boldsymbol{U}_{I_{j.}}^{\mathrm{T}}\boldsymbol{U}_{I_{j.}}\boldsymbol{V}\right)_{ik}}, \tag{5}$$

where $\boldsymbol{U}_{I_{j.}}$, and $\boldsymbol{R}_{I_{j.}}$ are submatrices of $\boldsymbol{U}$ and $\boldsymbol{R}$ with row indices in $I_{j.}$ The multiplication of the resulting $\boldsymbol{U}$ and $\boldsymbol{V}$ factor matrices produce a complete item response matrix $\boldsymbol{R}$. The maximum and minimum values of the imputed values are constrained to the highest and lowest possible responses to avoid generating out-of-range values.

Illustration of the NNMF method

For illustration purposes, we first consider a fully observed item response matrix (Supplementary Fig. 1) for five respondents and five items having five response categories with numeric values from 1 to 5.

	Respondents				
	1	2	3	4	5
Item 1	4	2	4	5	3
Item 2	2	3	2	3	4
Item 3	2	5	2	1	1
Item 4	4	2	4	5	3
Item 5	2	4	2	5	5

Supplementary Figure 2-0-1. Fully observed item response matrix.

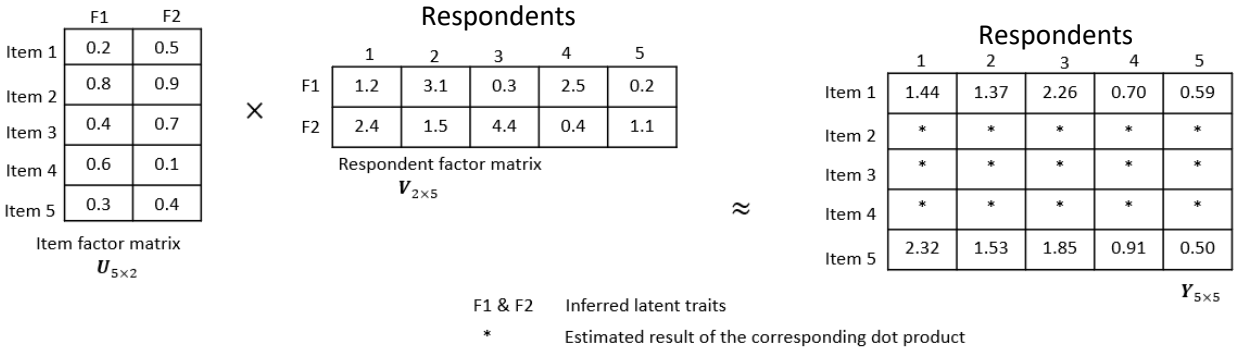
From Supplementary Fig. 1, we see that Respondents 1 and 3 have similar values on all the items. Responses on Items 1 and 4 are similar across all respondents. There are other relationships amongst respondents and items, and we use these associations to predict the missing values.

	Respondents				
	1	2	3	4	5
Item 1	4	2	4	5	3
Item 2	2	3	2	3	4
Item 3	2	5	2	1	1
Item 4	4	2	4	5	3
Item 5	2	4	M	5	5

Supplementary Figure 2-0-2. Item response matrix with missing response (M) on Item 5.

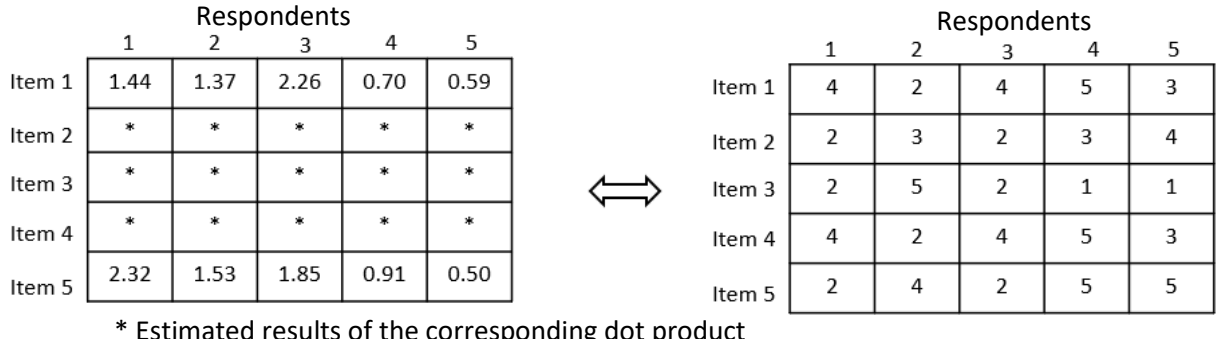
Consider the case where Respondent 3 has a missing response for Item 5 (Supplementary Fig. 2). The predicted missing value is 2 based on the association of the responses for Respondents 1 and 3. The missing value is straightforward to predict when the sample size and number of items are small. With large sample sizes and number of items optimization techniques are used to predict the missing values. The steps in the optimization process are:

Step 1: Randomly initialize the two factor matrices, item factor matrix $\boldsymbol{U}$ and respondent factor matrix $\boldsymbol{V}$ with two latent traits and compute the dot product. In Supplementary Fig. 3, **Y** contains the dot product of the two factor matrices. The first entry was obtained from the product of the first row in $\boldsymbol{U}$ and the first column in $\boldsymbol{V}$ and summing these values, that is [(0.2 x 1.2) + (0.5 x 2.4)] = 1.44. Other values in **Y** are calculated in the same manner.



Supplementary Figure 2-0-3. Initial factor matrices and their dot product (**Y**).

Step 2: Compare the output from Step 1 (Supplementary Fig. 3) to the actual item response matrix (Supplementary Fig. 1).



Supplementary Figure 2-0-4. Estimated (left) and actual (right) response matrices.

The sum of squares of the differences for each pair of responses in the estimated and actual response matrices (i.e., $[(4 - 1.44)^2 + (2 – 1.37)^2 + … + (5 – 0.50)^2]$) is used to evaluate how far or close the estimated item response matrix is from the actual response matrix. This is repeated until the responses from the estimated and actual response matrices are close.

Now, consider the case where the response matrix is not fully observed (Supplementary Fig. 5) because of missing responses on one or more items. Following these two steps, we can predict the missing responses in Supplementary Fig. 5. However, in this case, the sum of squares of the differences for each pair of responses in the estimated and actual response matrices is based on the observed responses only.

	Respondents				
	1	2	3	4	5
Item 1	4	M	4	5	M
Item 2	2	3	2	3	4
Item 3	2	M	2	M	1
Item 4	M	2	M	5	M
Item 5	2	4	M	5	M

Supplementary Figure 2-0-5. Item response matrix with missing responses (M).

Online Resource 2: Methods

Simulation study data generation

We simulated two-dimensional polytomous data for 30 items using the multidimensional graded response model (MGRM) [32, 33]:

$$P\left(\boldsymbol{R}_{ij} \geq c \mid \boldsymbol{\theta}, \alpha_j, \beta_{jc}\right) = \frac{exp\left[\sum_{m=1}^{M} \alpha_{jm}\left(\theta_{im} - \beta_{jc}\right)\right]}{1 + exp\left[\sum_{m=1}^{M} \alpha_{jm}\left(\theta_{im} - \beta_{jc}\right)\right]}, \quad (6)$$

where $P\left(\boldsymbol{R}_{ij} \geq c \mid \boldsymbol{\theta}, \alpha_j, \beta_{jc}\right)$ is the probability of selecting the response option c $(c = 1, \cdots, C)$ or higher on item j $(j = 1, \cdots, p)$ by respondent i $(i = 1, \cdots, n)$, with latent trait $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_M)$, α_{jm} is the slope or discrimination parameter of item j on dimension m, and β_{jc} is the threshold or difficulty of item j for the c^{th} category.

Statistical analyses of Joint Replacement Registry data

The dimensionality of the SF-12 items was assessed using exploratory and confirmatory analyses, as described in previous studies [65, 66]. The exploratory approach compared the first and second eigenvalues from the polychoric correlations for all the items. The first eigenvalue should be significantly greater than the second eigenvalue (i.e., ratio > 4:1) to support unidimensionality [65]. The confirmatory approach compared the comparative fit index (CFI), root mean square error of approximation (RMSEA), and standardized root mean squared residual (SRMR) of unidimensional and two-dimensional graded response models using a unit variance identification constraint (i.e., constraining the mean and variance of each factor to 0 and 1, respectively). The following criteria were used to evaluate model fit: RMSEA ≤ 0.05, SRMR ≤ 0.08, CFI > 0.95 indicate close or good model fit [67, 68]. Items with complete responses was used to assess the dimensionality of the SF-12.

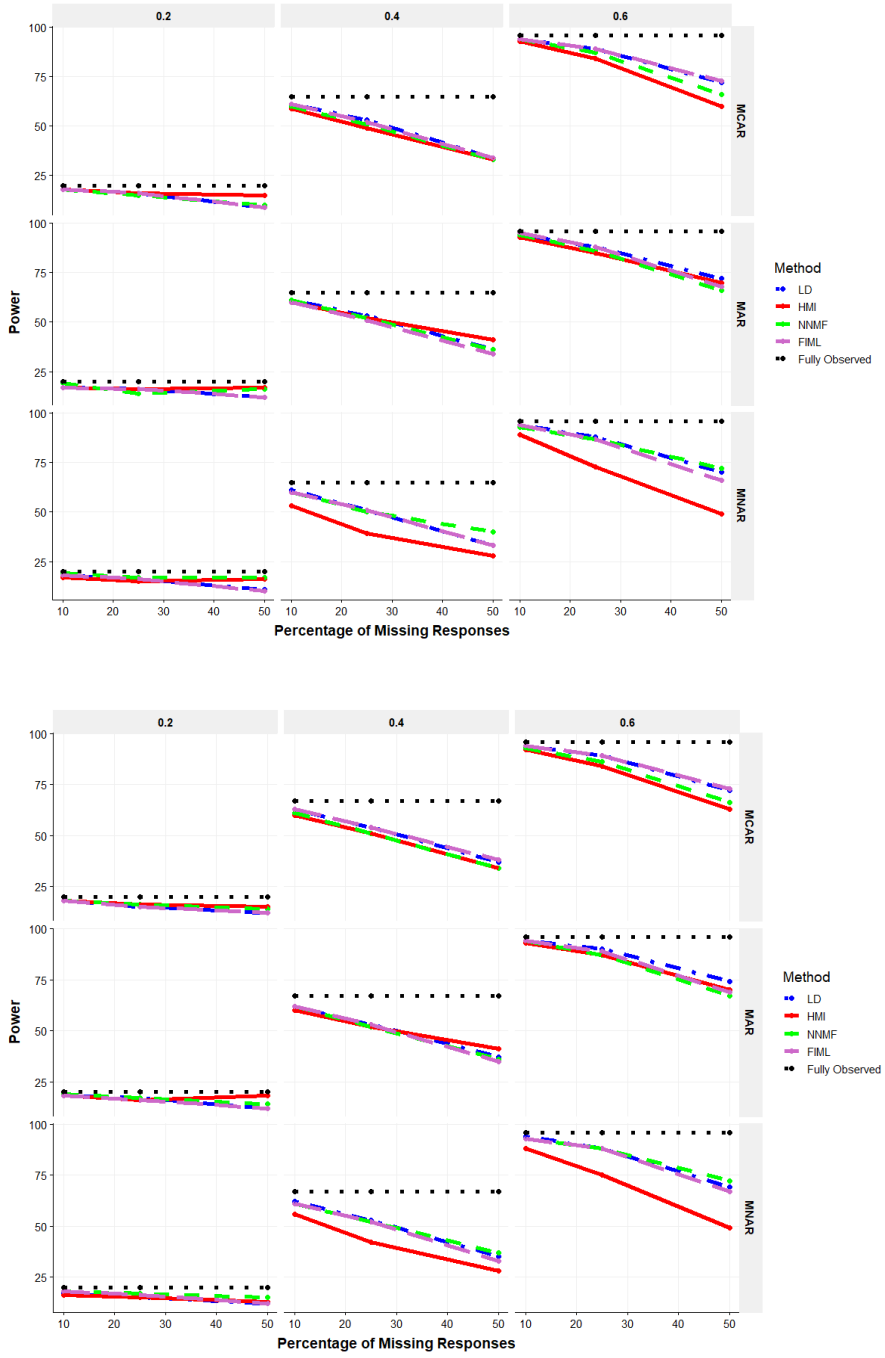
At a minimum, one anchor item is needed for each dimension to define the latent trait on which the groups are compared. We adopted a method described previously [69] to select one anchor item for each dimension. First, we fit a model (i.e., the model with the best fit statistics) to the data in which the slopes and thresholds of all the items were constrained equal across the two groups (i.e., age groups, males/females). Second, a series of unconstrained models were fitted to the data using the all-other anchor approach (AOAA), which uses all items except for the current item as the anchor. A LRT was used to test the difference between the constrained and unconstrained models for each item. We labelled an item as DIF-free if the LRT statistic was not statistically significant. Next, we selected the DIF-free items that had the largest slope parameters as the anchor items for the dimension(s). This approach has been shown to outperform other anchor selection strategies [57, 58].

For the assessment of uniform and non-uniform DIF on the items of the PCS and MCS, we specified an unconstrained DIF model, where the slopes and thresholds of all items, except for the anchor items, were freely estimated. Next, we fitted a set of constrained models (i.e., no-DIF models), where the parameters of the remaining items were constrained one at a time. The LRT statistic was used to test the difference between the unconstrained DIF and no-DIF models. A statistically significant LRT statistic when an unconstrained DIF model was compared to: (1) a no-DIF model (i.e., the threshold parameters were constrained for the two groups), indicates the presence of uniform DIF, and (2) a no-DIF model, where the slope parameters are constrained for the two groups, indicates the presence of non-uniform DIF.

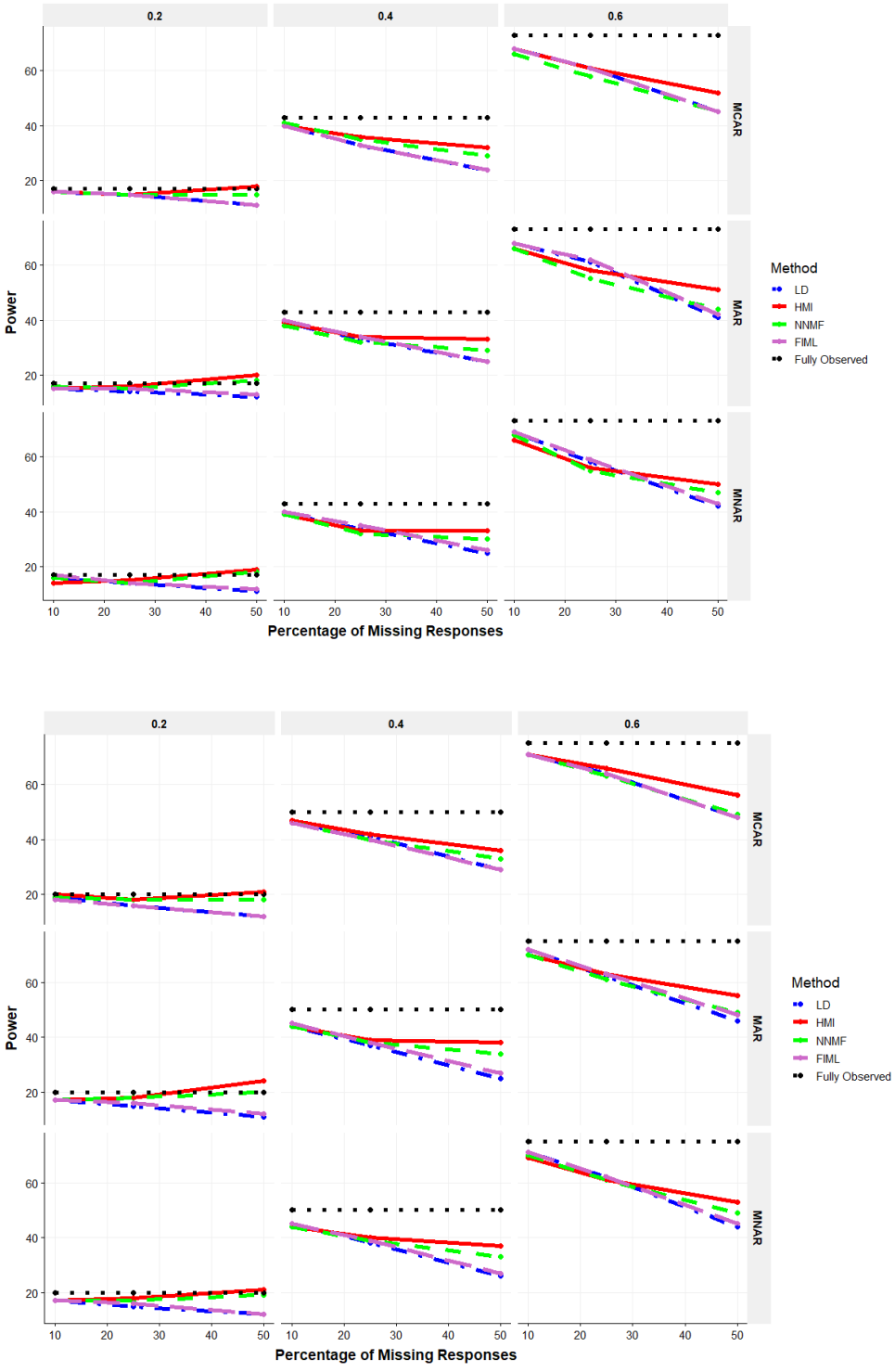
Model parameters were estimated using the maximum likelihood method. We used the Hochberg step-up approach to control for the family wise Type I error rate to $\alpha = 0.05$. The Hochberg approach defines the significance (α_t) as $\alpha/m\text{-}t\text{+}1$, where m is the total number of hypotheses

tested and $t = m, m\text{-}1, \ldots, 1$. This approach arranges the *p*-values from largest to smallest and the corresponding hypotheses to be tested. Hypothesis testing starts with the item that is associated with the largest *p*-value. All hypotheses are rejected if the largest *p*-value is less than α. If the hypothesis associated with the largest *p*-value is not rejected, the next largest *p*-value is compared to α/2. If this next-largest *p*-value is less than α/2, the corresponding hypothesis and all subsequent hypotheses are rejected. This process continues until a statistically significant result is obtained and the corresponding hypothesis is rejected along with all subsequent hypotheses, or until all *p*-values have been tested.

Online Resource 3: Power to detect uniform and non-uniform DIF



Supplementary Figure 2-0-6. Power (%) of the LRT to detect **uniform DIF** for latent trait correlations of 0.3 (top panel), and 0.5 (bottom panel) and sample size of R500/F500 (R: reference group and F: focal group) stratified by the missing data mechanism and magnitude of DIF (20% of the items exhibited DIF). MCAR: missing completely at random; MAR: missing at random; and MNAR: missing not at random.



Supplementary Figure 2-0-7. Power (%) of the LRT to detect **non-uniform DIF** for latent trait correlations of 0.3 (top panel), and 0.5 (bottom panel) and sample size of R500/F500 (R: reference group and F: focal group) stratified by the missing data mechanism and magnitude of DIF (20% of the items

exhibited DIF). MCAR: missing completely at random; MAR: missing at random; and MNAR: missing not at random.

Online Resource 4: Power to detect uniform and non-uniform DIF for sample size R250/F250

Supplementary Table 2-1. Power (%) of the LRT to detect uniform and non-uniform DIF for sample size R250/F250 (R: reference group and F: focal group) and magnitude of DIF (20% of the items exhibited DIF effect of 0.6) stratified by missing data mechanism and latent trait correlations

DIF	Mech.	ρ	Percentage of missing responses											
			10%				25%				50%			
			LD	HMI	NNMF	FIML	LD	HMI	NNMF	FIML	LD	HMI	NNMF	FIML
Uniform	MCAR	0.3	67.4	65.0	65.9	67.5	57.6	52.8	54.7	58.1	39.2	35.7	36.5	40.1
		0.5	66.4	65.6	66.3	66.2	60.5	55.3	57.4	61.3	40.8	38.6	38.2	40.9
	MAR	0.3	66.1	64.9	66.0	66.4	57.6	55.1	56.5	56.9	39.4	41.5	35.6	37.7
		0.5	67.2	65.6	66.0	67.2	57.9	57.0	56.5	58.2	40.7	43.7	37.4	37.4
	MNAR	0.3	65.1	58.2	64.9	65.2	55.0	40.8	53.3	54.7	36.6	26.3	39.4	34.9
		0.5	67.1	59.8	66.2	68.1	56.7	44.5	55.8	56.5	38.5	30.0	42.2	37.1
Non-uniform	MCAR	0.3	43.2	43.7	42.9	43.5	38.0	37.9	36.4	37.5	28.1	37.4	30.1	27.9
		0.5	45.8	46.7	45.6	46.0	40.8	42.8	42.1	40.7	27.8	37.8	31.2	27.5
	MAR	0.3	43.4	41.9	42.3	43.7	37.1	37.8	35.0	37.6	25.2	34.3	28.5	26.0
		0.5	46.7	44.9	44.0	46.9	37.0	36.9	35.1	37.2	25.8	37.4	32.2	26.7
	MNAR	0.3	44.6	41.9	43.6	44.6	37.1	37.2	36.1	37.8	24.4	34.3	30.4	25.6
		0.5	45.2	44.2	44.9	45.6	37.8	39.0	37.8	38.5	25.9	36.9	33.0	26.4

MCAR: Missing completely at random; MAR: Missing at random; MNAR: Missing not at random; LD: List-wise deletion; HMI: Half-mean imputation; NNMF: Non-negative matrix factorization; FIML: full information maximum likelihood; ρ: latent trait correlation

Online Resource 5: Differences in the log-likelihood statistics of the unconstrained and constrained DIF models by age group

Supplementary Table 2-2. Difference in the log-likelihood statistics (degrees of freedom) of the unconstrained and constrained DIF models by age group for the SF-12 physical component summary (PCS) and mental component summary (MCS) items

Item	**Uniform DIF**				**Non-uniform DIF**			
Physical Component Summary (PCS)	**LD**	**HMI**	**NNMF**	**FIML**	**LD**	**HMI**	**NNMF**	**FIML**
PCS1: General health	**23.90 (4)**	**26.45 (4)**	**18.22 (4)**	**24.22 (4)**	0.99 (1)	0.86 (1)	2.68 (1)	0.91 (1)
PCS2: Limits in moderate activity	6.75 (2)	2.56 (2)	2.70 (2)	3.98 (2)	0.02 (1)	0.06 (1)	0.00 (1)	0.06 (1)
PCS3: Climbing several flights	4.92 (2)	1.85 (2)	1.41 (2)	2.53 (2)	0.00 (1)	0.32 (1)	0.27 (1)	0.37 (1)
PCS4: Accomplished less	†	†	†	†	†	†	†	†
PCS5: Limited in work and other activities	6.45 (4)	7.13 (4)	6.00 (4)	7.20 (4)	5.08 (1)	5.34 (1)	5.80 (1)	5.64 (1)
PCS6: Pain interference with normal work	5.10 (4)	4.77 (4)	5.77 (4)	4.47 (4)	4.70 (1)	4.76 (1)	4.92 (1)	3.47 (1)
Mental Component Summary (MCS)								
MCS1: Accomplished less	†	†	†	†	†	†	†	†
MCS2: Less careful than usual	11.14 (4)	14.10 (4)	**15.21 (4)**	**15.45 (4)**	5.99 (1)	6.75 (1)	4.45 (1)	**8.84 (1)**
MCS3: Felt calm and peaceful	**25.95 (4)**	**26.96 (4)**	13.92 (4)	**24.23 (4)**	1.59 (1)	0.17 (1)	1.57 (1)	0.77 (1)
MCS4: Have a lot of energy	7.66 (4)	5.99 (4)	8.88 (4)	7.46 (4)	2.49 (1)	0.54 (1)	0.54 (1)	1.49 (1)
MCS5: Felt downhearted and depressed	10.20 (4)	8.73 (4)	3.28 (4)	8.27 (4)	0.67 (1)	0.01 (1)	2.80 (1)	0.21 (1)
MCS6: Emotional problems	1.82 (4)	1.06 (4)	0.52 (4)	1.36 (4)	3.74 (1)	2.26 (1)	0.01 (1)	4.05 (1)

†= anchor item; Boldface font indicates a statistically significant p-value; LD: List-wise deletion; HMI: Half-mean imputation; NNMF: Non-negative matrix factorization; FIML: full information maximum likelihood

Online Resource 6: Sensitivity analysis to assess the impact of unequal sample sizes on Type I error rates

Supplementary Table 2-3. Type I error rates (%) for fully observed data and item non-response methods when testing for DIF using the likelihood ratio test when none of the 20% target items exhibited DIF effects for R1500/F500 (R: reference group and F: focal group) and correlation between latent traits

Mech.	ρ	Fully Observed	Percentage of missing responses											
			10%				25%				50%			
			LD	HMI	NNMF	FIML	LD	HMI	NNMF	FIML	LD	HMI	NNMF	FIML
MCAR	0.3	5.9	5.4	5.4	5.7	5.4	5.3	6.0	5.6	5.4	5.3	**9.3**	7.0	5.2
	0.5	5.2	5.5	5.4	5.6	5.5	5.4	6.1	6.7	5.2	5.0	**10.5**	7.4	5.1
MAR	0.3	5.9	5.9	6.3	5.4	6.0	5.7	**7.6**	6.3	5.8	5.3	**10.2**	7.2	5.4
	0.5	5.2	5.4	5.8	5.8	5.4	5.4	6.4	6.0	5.3	5.2	**10.8**	**7.6**	5.0
MNAR	0.3	5.9	6.0	6.2	5.8	6.0	5.8	**7.8**	6.5	5.8	5.7	**9.6**	**8.0**	5.8
	0.5	5.2	5.4	5.5	5.4	5.4	5.3	6.3	6.2	5.5	5.1	**10.7**	7.2	5.2

MCAR: Missing completely at random; MAR: Missing at random; MNAR: Missing not at random; LD: List-wise deletion; HMI: Half-mean imputation; NNMF: Non-negative matrix factorization; FIML: full information maximum likelihood; Boldface font indicates error rates that fall outside the bounds of Bradley's [49] liberal criterion (i.e., 2.5 – 7.5%); ρ: latent trait correlation

Online Resource 7: Sensitivity analysis to assess the impact of unequal sample sizes on statistical power

Supplementary Table 2-4. Power (%) of the LRT to detect uniform and non-uniform DIF for sample size of R1500/F500 (R: reference group and F: focal group), and magnitude of DIF (20% target items exhibited DIF effect of 0.2) stratified by missing data mechanism

DIF	Mech.	ρ	Fully Observed	Percentage of missing responses											
				10%				25%				50%			
				LD	HMI	NNMF	FIML	LD	HMI	NNMF	FIML	LD	HMI	NNMF	FIML
Uniform	MCAR	0.3	27.7	24.3	23.5	23.6	24.5	19.4	19.9	19.7	19.6	13.9	18.5	15.2	14.4
		0.5	27.8	26.1	24.8	25.1	25.9	22.1	20.2	21.8	22.1	15.4	17.9	17.8	15.5
	MAR	0.3	27.7	25.4	25.8	24.7	25.6	20.6	22.4	21.5	20.4	15.3	21.9	17.0	15.3
		0.5	27.8	25.4	26.0	25.6	25.5	22.2	22.5	22.4	21.3	14.9	20.2	17.2	13.8
	MNAR	0.3	27.7	23.4	21.5	22.9	23.3	17.5	15.4	18.9	17.8	13.5	15.7	16.8	12.8
		0.5	27.8	23.6	22.1	23.6	23.8	18.6	17.0	20.0	18.2	13.3	18.1	18.3	12.9
Non-uniform	MCAR	0.3	21.7	20.5	20.7	20.8	20.4	17.9	19.1	18.9	18.1	15.3	22.2	18.1	14.9
		0.5	28.1	27.0	26.9	27.5	27.2	23.2	26.8	25.9	23.5	17.7	27.4	23.2	18.1
	MAR	0.3	21.7	19.5	19.7	18.6	19.8	16.1	19.0	17.8	17.2	12.8	21.1	18.1	13.2
		0.5	28.1	25.6	24.9	25.0	25.9	21.1	23.4	22.7	21.5	14.5	27.3	22.5	14.9
	MNAR	0.3	21.7	19.3	20.0	20.1	19.6	16.2	19.4	19.0	16.5	11.5	20.4	18.7	12.5
		0.5	28.1	24.9	24.9	25.5	24.8	20.2	24.0	24.2	20.2	14.7	27.4	24.0	15.2

MCAR: Missing completely at random; MAR: Missing at random; MNAR: Missing not at random; LD: List-wise deletion; HMI: Half-mean imputation; NNMF: Non-negative matrix factorization; FIML: full information maximum likelihood; R: Reference group; F: Focal group; ρ: latent trait correlation

Supplementary Table 2-5. Power (%) of the LRT to detect uniform and non-uniform DIF for sample size of R1500/F500 (R: reference group and F: focal group), and magnitude of DIF (20% target items exhibited DIF effect of 0.4) stratified by missing data mechanism

DIF	Mech.	ρ	Fully Observed	Percentage of missing responses											
				10%				25%				50%			
				LD	HMI	NNMF	FIML	LD	HMI	NNMF	FIML	LD	HMI	NNMF	FIML
Uniform	MCAR	0.3	85.1	80.4	78.9	79.5	80.4	71.9	67.5	69.9	71.9	51.5	46.8	47.5	51.5
		0.5	85.9	82.0	79.7	81.4	82.3	72.0	66.3	69.9	72.0	53.7	47.1	49.5	54.1
	MAR	0.3	85.1	80.7	79.9	79.6	80.5	71.8	68.9	69.0	70.6	52.2	53.6	49.1	49.4
		0.5	85.9	81.7	79.8	81.0	81.9	72.7	70.4	71.1	72.4	50.8	53.6	49.5	47.3
	MNAR	0.3	85.1	77.9	70.4	75.5	77.5	65.3	50.8	64.6	64.5	44.2	34.0	46.8	41.4
		0.5	85.9	79.8	72.5	78.3	79.1	66.5	52.8	65.9	64.9	42.9	36.7	49.2	40.7
Non-uniform	MCAR	0.3	60.3	56.3	56.0	55.9	56.6	49.2	49.3	47.6	49.2	37.8	42.0	36.8	37.2
		0.5	65.3	61.2	62.5	61.2	61.2	54.2	57.4	55.4	54.3	41.4	49.5	43.6	41.2
	MAR	0.3	60.3	55.0	54.4	52.7	55.7	46.3	46.4	44.7	47.5	32.8	41.5	35.4	33.6
		0.5	65.3	60.1	59.7	58.9	60.5	52.3	52.4	50.7	53.0	36.4	48.7	42.5	36.6
	MNAR	0.3	60.3	54.0	53.1	54.1	54.3	45.5	46.3	46.7	46.1	32.3	38.7	34.3	32.9
		0.5	65.3	59.7	58.1	60.0	60.1	50.0	52.1	52.6	50.4	36.0	48.1	44.1	35.8

MCAR: Missing completely at random; MAR: Missing at random; MNAR: Missing not at random; LD: List-wise deletion; HMI: Half-mean imputation; NNMF: Non-negative matrix factorization; FIML: full information maximum likelihood; R: Reference group; F: Focal group; ρ: latent trait correlation

Supplementary Table 2-6. Power (%) of the LRT to detect uniform and non-uniform DIF for sample size of R1500/F500 (R: reference group and F: focal group), and magnitude of DIF (20% target items exhibited DIF effect of 0.6) stratified by missing data mechanism

DIF	Mech.	ρ	Fully Observed	Percentage of missing responses											
				10%				25%				50%			
				LD	HMI	NNMF	FIML	LD	HMI	NNMF	FIML	LD	HMI	NNMF	FIML
Uniform	MCAR	0.3	99.5	99.2	98.9	99.1	99.2	97.5	95.9	97.2	97.5	88.7	80.6	84.5	88.9
		0.5	99.6	98.9	98.7	98.8	98.9	97.6	95.5	96.6	97.7	89.2	79.7	84.5	89.3
	MAR	0.3	99.5	98.9	98.9	99.0	99.0	97.9	96.7	97.1	97.7	88.9	86.5	85.6	86.7
		0.5	99.6	99.2	98.8	98.9	99.3	98.0	96.6	97.5	97.8	87.8	85.7	84.4	84.4
	MNAR	0.3	99.5	98.8	96.7	98.4	98.8	95.7	84.7	95.3	95.6	81.1	60.5	82.8	78.9
		0.5	99.6	98.9	97.1	99.0	99.0	96.4	86.3	95.6	95.7	80.1	62.8	83.6	77.0
Non-uniform	MCAR	0.3	85.6	82.7	81.6	81.2	82.5	76.1	75.8	74.0	76.3	61.5	63.8	56.2	61.6
		0.5	87.7	84.7	84.7	84.2	84.6	79.5	80.2	78.2	79.7	66.0	69.8	63.4	65.5
	MAR	0.3	85.6	81.1	80.1	79.1	81.4	74.6	72.4	69.9	74.8	57.2	62.7	54.2	58.1
		0.5	87.7	84.4	82.5	82.1	84.5	77.2	77.5	73.4	77.8	58.5	68.1	59.5	59.3
	MNAR	0.3	85.6	80.9	78.6	79.6	81.6	72.3	72.3	71.6	73.1	56.5	58.9	53.6	56.6
		0.5	87.7	83.5	81.1	82.3	83.8	76.2	76.2	76.3	76.8	59.4	67.0	62.4	59.1

MCAR: Missing completely at random; MAR: Missing at random; MNAR: Missing not at random; LD: List-wise deletion; HMI: Half-mean imputation; NNMF: Non-negative matrix factorization; FIML: full information maximum likelihood; R: Reference group; F: Focal group; ρ: latent trait correlation

CHAPTER 3: ESTIMATING LONGITUDINAL CHANGE IN LATENT VARIABLE MEANS: A COMPARISON OF NON-NEGATIVE MATRIX FACTORIZATION AND OTHER ITEM NON-RESPONSE METHODS

3.1 Overview

This study extends the methods from the previous chapter to a longitudinal design. Longitudinal change in latent variables, such as health-related-quality-of-life, is often of interest in clinical studies to evaluate the effectiveness of interventions. However, estimates of change in latent variable, including means are affected by item non-response. In this study, I assess the performance of non-negative matrix factorization, multiple imputation with conditional proportional odds model, full information maximum likelihood, and complete-case analysis when estimating change in latent variable means over time. This objective was conducted in the context of longitudinal latent variable measurement models under the assumption of longitudinal measurement invariance (i.e., equivalence), which implies that the latent variable carries the same meaning and the same scale over all measurement occasions, and partial measurement invariance. Monte Carlo simulation and numeric example were used to evaluate the bias and mean squared error of the methods. The non-negative matrix factorization method outperformed the proportional odds model and full information maximum likelihood methods when the latent variable correlations over time were strong and percentage of missing data was 25% or more.

Publication Details

Ayilara, O.F., Sajobi, T.T., Barclay, R., Bohm, E., Jafari Jozani, M., & Lix, L.M. (2022). Estimating longitudinal change in latent variable means: a comparison of non-negative matrix factorization and other item non-response methods. *Journal of Statistical Computation and Simulation,*1-20.

3.2 Abstract

Estimates of longitudinal change in the parameters of latent (i.e., unobserved) variables, including means, are affected by non-response on the items or indicators of the latent variable. This study used Monte Carlo simulation and a numeric example to compare four ordinal item non-response methods: non-negative matrix factorization (NNMF), multiple imputation with conditional proportional odds model (POM), full information maximum likelihood (FIML), and complete-case analysis, when estimating longitudinal change in latent variable means. The mean squared error for the NNMF method was more than 40% lower than for the FIML and POM methods when the latent variable correlations over time were strong, percentage of missing data was 25% or more, and sample size was large. The NNMF method is a promising method to address item non-response. It is relatively efficient when sample size is large, and the percentage of missing data is high, but has limitations under other data-analytic conditions.

3.3 Introduction

Longitudinal change in latent (i.e., unobserved) variables such as health-related quality of life and other aspects of well-being, are used to describe disease trajectories [1] and evaluate the effectiveness of interventions [2]. Latent variables are often measured by a set of related items or indicators (i.e., observed variables) that reflect the underlying construct. The response options of the items are often ordinal (e.g., Likert scales). Longitudinal change in latent variable means can be estimated using classical test theory (CTT) or item response theory (IRT) models. CTT relates individual's observed score (i.e., unweighted sum of the individual's responses to items) to his or her location on the latent variable, while IRT uses latent characterizations of individuals and items to explain the observed responses [3–6]. CTT and IRT differ in their assumptions when estimating longitudinal change in latent variable means. However, responses at the item-level form the basis of analyses for both theories.

Estimates of change in latent variable parameters, including change in latent variable means are affected by non-response (i.e., missing data) on the items used to measure the variable. Item non-response may occur because participants miss scheduled clinic visits, have difficulty understanding or interpreting the questions, are uncomfortable answering the questions or dropout from the study due to death or migration. The potential consequences of item non-response include a loss of statistical power to detect longitudinal change in latent variables, and over- or under-estimation of the magnitude of change [7, 8].

Item non-response may be challenging to address because of its ordinal nature. Previous studies about item non-response methods have primarily relied on the assumption that the data follow a normal distribution [9–11]. This assumption is likely to be violated, particularly when the number of response options for the items are small [12]. The choice and performance of methods

to address item non-response is dependent on the type of variable with the missing response and missingness mechanism.

Using Little and Rubin's taxonomy, missingness can be categorized as missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR) [13]. Suppose we have a simple random sample of n individuals from the population of interest. Let $\boldsymbol{Y}_i = \left(Y_{i1}, Y_{i2}, \ldots, Y_{ip}\right)^{\mathrm{T}}$ denote the vector of p item responses, $\boldsymbol{Y}_i^O$ denote the subset of p items that are observed, and $\boldsymbol{Y}_i^M$ denote the subset of p items for which responses are missing for the i^{th} individual ($i = 1, \ldots, n$). Let $\delta_{ij} = 1$ if Y_{ij} is observed, and $\delta_{ij} = 0$ if Y_{ij} is missing for $j = 1, \ldots, p$ and $\boldsymbol{\delta}_i = \left(\delta_{i1}, \ldots, \delta_{ip}\right)^{\mathrm{T}}$. Data are MCAR if the probability of missingness is unrelated to the observed and unobserved outcomes, that is, $P(\boldsymbol{\delta}_i|\boldsymbol{Y}_i) = P(\boldsymbol{\delta}_i)$. Data are MAR, if conditional on the observed data, the probability of missingness is independent of the unobserved outcomes, i.e., $P(\boldsymbol{\delta}_i|\boldsymbol{Y}_i) = P(\boldsymbol{\delta}_i|\boldsymbol{Y}_i^O)$. Finally, data are MNAR if the probability of missing observations depends in whole or in part, on unobserved measurements, i.e., $P(\boldsymbol{\delta}_i|\boldsymbol{Y}_i) = P\left(\boldsymbol{\delta}_i\middle|\boldsymbol{Y}_i^O, \boldsymbol{Y}_i^M\right)$.

Commonly used methods to address ordinal item non-response are complete-case analysis (CCA) [14] and half-mean imputation (HMI) [7]. The CCA removes all individuals who have at least one item with a missing response, while the HMI method fills in missing item responses with the mean of the available item responses if 50% or more of the item responses are complete and removes individuals with more than 50% missing responses on all the items [15, 16]. These methods are easy to implement and are widely available in most statistical software packages such as SPSS, SAS and R. However, they may result in biased parameter estimates, especially when the percentage of non-response is high [16].

Multiple imputation (MI) and full information maximum likelihood (FIML) are other methods to address item non-response [9, 10, 16–24]. These two methods have been shown to perform equally well when the imputation model (i.e., model for generating the imputed values) and the analysis or estimation model (i.e., target model to address research objectives) are similar [22–24]. However, there are a number of potential issues in order to obtain valid parameter estimates when using the MI method to address item non-response. First, the functional form of the imputation model and the variables to include in the model need to be specified correctly [19, 25]. Second, any potential non-linear relationships that may exist amongst variables needs to be incorporated in the imputation model [26]. The FIML method rests on the assumption of correct model specification, as well as additional assumptions about the parametric probability model [7, 27]; when these assumptions are not tenable, the FIML method may give less than optimal performance [6].

Non-negative matrix factorization (NNMF), which is an unsupervised machine learning method that uses a set of optimization techniques to identify a low-dimensional data structure from the original data, can be used to address item non-response. NNMF has been previously used in machine learning for dimension reduction [28, 29], and feature extraction [30, 31] from high-dimensional non-negative data. The NNMF method has been used in recommender systems to predict movie ratings from an incomplete rating matrix [32–34]. The problem of predicting movie ratings can be reformulated as one of missing data imputation via matrix completion. Thus, this method is a plausible approach to address item non-response in non-negative high-dimensional data, that are discrete (i.e., ordinal) and highly correlated, which are the characteristics of item-level data used to measure latent variables.

The objective of our study was to compare the performance of the NNMF method with CCA, FIML and MI with conditional proportional odds model (POM), when estimating longitudinal change in latent variable means. We did this in the context of longitudinal latent variable measurement models under the assumption of longitudinal measurement invariance (i.e., equivalence), which implies that the latent variable carries the same meaning and the same scale over all measurement occasions, and partial measurement invariance [35, 36].

3.4 Methods

3.4.1 Item Non-Response Methods

In this section, we introduce NNMF and describe the MI and FIML methods [24].

NNMF

The NNMF method approximately decomposes a non-negative matrix into the product of two low-dimensional matrices [30]. It is assumed that there are correlations and underlying structures in the dataset such that only a subset of observed values is required to impute some or all missing values. The main difference between the NNMF method and other matrix factorization methods, such as singular value decomposition and maximum-margin matrix factorization, is that the two low-dimensional matrices are constrained to be non-negative. These constraints make the NNMF method applicable to address ordinal item non-responses [37].

Suppose we have a fully observed response matrix $\boldsymbol{Y}_{p\times n}$ with low rank k $[k \ll \min(n,p)]$, where p is the number of items or indicators, and n is the number of respondents. The goal of the NNMF method is to obtain an item factor matrix $\boldsymbol{U}_{p\times k}$, and a respondent factor matrix $\boldsymbol{V}_{k\times n}$, such that $\boldsymbol{Y} \approx \boldsymbol{UV}$, constraining $\boldsymbol{U}$ and $\boldsymbol{V}$ to be non-negative [30, 31]. The item factor matrix

depicts the relationship of each item to the latent variable, and the respondent factor matrix describes the relationship of each respondent to the latent variable.

The NNMF method involves minimizing:

$$P(\boldsymbol{U},\boldsymbol{V}) = \frac{1}{2} \|\boldsymbol{Y} - \boldsymbol{UV}\|^2, \quad (1)$$

subject to $\boldsymbol{U} \geq 0, \boldsymbol{V} \geq 0$, where $\|.\|^2$ is the squared Frobenius norm of the matrix [29]. We used multiplicative update rules, an iterative method to update factor matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ [38], because of their ease of implementation and speed of convergence. The multiplicative update rules for entries u_{ik} and v_{kj} of the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ respectively are:

$$u_{ik} \Leftarrow \frac{(\boldsymbol{YV}^{\mathrm{T}})_{ik} u_{ik}}{(\boldsymbol{UVV}^{\mathrm{T}})_{ik}}, \quad (2)$$

$$v_{kj} \Leftarrow \frac{(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{Y})_{kj} v_{kj}}{(\boldsymbol{U}^{T}\boldsymbol{U}\,\boldsymbol{V})_{kj}}, \quad (3)$$

Now, consider the situation where some entries in $\boldsymbol{Y}$ are missing. The initial formulation of the multiplicative update rules will be insufficient unless additional constraints are imposed on the rows and columns of $\boldsymbol{U}$ and $\boldsymbol{V}$. For an incomplete ordinal item response matrix $\boldsymbol{Y}$ with low rank k, let $\omega \subset \{1, \ldots, p\} \times \{1, \ldots, n\}$ denote the indices of observed entries, $I_j = \{i: y_{ij} \text{ is observed}\}$ and $\tilde{I}_j = \{i: y_{ij} \text{ is missing}\}$. Since $\boldsymbol{Y}$ is assumed to have a low rank k, incomplete item responses in $\boldsymbol{Y}$ can be used to achieve the approximation $\boldsymbol{Y} \approx \boldsymbol{UV}$. The optimization problem in Eq.(1) can then be reformulated as,

$$\underset{U,V}{\text{minimize}} \ \frac{1}{2} \sum_{(i,j)\in\omega} (\boldsymbol{Y}_{ij} - (\boldsymbol{UV})_{ij})^2, \quad (4)$$

subject to $\boldsymbol{U} \geq 0, \boldsymbol{V} \geq 0$. When applying the multiplicative update rules on the j^{th} column of $\boldsymbol{V}$ in Eq.(4), all $\tilde{I}_j$ rows of $\boldsymbol{U}$ are removed. The resulting multiplicative update equation with squared Frobenius norm becomes,

$$v_{kj} \Leftarrow \frac{\left(\boldsymbol{U}_{I_{j.}}^{\mathrm{T}} \boldsymbol{Y}_{I_{j.}}\right)_{kj} v_{kj}}{\left(\boldsymbol{U}_{I_{j.}}^{\mathrm{T}} \boldsymbol{U}_{I_{j.}} \boldsymbol{V}\right)_{ik}}, \tag{5}$$

where $\boldsymbol{U}_{I_{j.}}$, and $\boldsymbol{Y}_{I_{j.}}$ are submatrices of $\boldsymbol{U}$ and $\boldsymbol{Y}$ with row indices in $I_{j.}$ The multiplication of the resulting $\boldsymbol{U}$ and $\boldsymbol{V}$ factor matrices produce an approximation to the original data points in item response matrix $\boldsymbol{Y}$. The maximum and minimum imputed values are constrained to the highest and lowest possible item responses to avoid generation of out-of-range values.

MI

MI is a widely used method to address ordinal missing data that are MCAR or MAR. The MI method is implemented in three stages. The first stage, which is the *imputation stage*, involves specifying a parametric probability model for the missing responses given the observed data, and drawing missing responses from the posterior predictive distribution $M > 1$ times. The second stage, known as the *analysis stage*, involves fitting a substantive model to the M imputed datasets. Finally, in the *pooling stage*, the M parameter estimates are combined using Rubin's rules [39] to produce a final set of results.

When using the MI technique, it is common to treat ordinal response as if they were continuous and generate imputed datasets assuming a multivariate normal distribution. This approach has been shown to produce reliable estimates when imputing Likert-type ordinal missing values that were subsequently aggregated to produce scale scores [10]. However, this approach is not designed to address ordinal item non-response and may produce biased results when the

distributions of the ordinal variables are severely asymmetric [9, 10]. Given this reality, van Buuren [19] recommended using an appropriate method designed for ordinal responses. Hence, we use the fully conditional specification (FCS) with proportional odds model (POM) in our study. The FCS is an imputation method that specifies the multivariate MI model on a variable-by-variable basis using a set of conditional densities [40].

Let Y_j denote the j^{th} incomplete item, $\boldsymbol{Y}_{-j} = \left(Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_p\right)^{\mathrm{T}}$ denote all the variables in $\boldsymbol{Y}$ excluding Y_j, and $\boldsymbol{\delta}_{-j} = \left(\delta_1, \ldots, \delta_{j-1}, \delta_{j+1}, \ldots, \delta_p\right)^{\mathrm{T}}$ denote the set of $p-1$ missing items, where $\delta_j = 1$ if Y_j is observed, and 0 otherwise. The FCS method defines the multivariate distribution $P(\boldsymbol{Y}, \boldsymbol{\delta}|\boldsymbol{\vartheta})$ through a set of conditional densities $P(Y_j|\boldsymbol{Y}_{-j}, \boldsymbol{\delta}, \phi_j)$ for each Y_j, where $\boldsymbol{\vartheta}$ and ϕ_j are the vectors of unknown parameters for the multivariate and conditional densities, respectively. Starting with reasonable predicted or guessed values, the conditional density is used to impute $Y_j|\boldsymbol{Y}_{-j}, \boldsymbol{\delta}$. In this study, we used the chained equation algorithm because of its availability and flexibility to impute missing data on a variable-by-variable basis [40, 41].

FIML

The FIML or direct likelihood method has only recently become a widely-used approach to address ordinal non-response in item response models because of its availability in statistical software packages such as Mplus and R [42]. This method is similar to the maximum likelihood method, which chooses parameter values that assign the maximum possible probability or probability density to the observed data under a well-defined family of parametric probability models. Without loss of generality, the sample log-likelihood function can be written as:

$$l = \sum_{i=1}^{n} \ln\left(f(y_i)\right) \tag{6}$$

where $f(y_i)$ is the probability density function for the vector of observed variables in $\boldsymbol{Y}$ for the i^{th} individual. The likelihood function of the two-tier longitudinal graded response model (GRM) [43] has a complicated form that requires computational techniques such as the expectation-maximization (EM) algorithm [44]. Estimates obtained using the FIML method are unbiased if the missing data are MAR and the model has been correctly specified [7].

3.4.2 Simulation Study

Our simulation study examined the performance of the NNMF method and compared it with the CCA, FIML, and POM methods. We considered two simulation scenarios to reduce any potential risk of confounding our results with model specifications. In the first scenario, which assumes full invariance (i.e., equivalence) of the latent variable measurement model over time, the slopes and thresholds of the items in the population model (i.e., a two-tier longitudinal graded response single-factor model [43], depicted in Figure 1), were constrained to be equal across both measurement occasions [45]. In the second scenario, which assumes partial invariance, the slopes and thresholds for 50% of the items in the population model in Figure 1 were not equal at the first and second measurement occasions [11, 46]. The magnitude of inequality between the first and second measurement occasions was specified to be either moderate or large for both the slope and threshold parameters. The former condition of inequality was defined as 0.3 and the latter was defined as 0.6, respectively [11].

In both scenarios, the number of items was fixed to eight, the variances of the latent variable were fixed to one, the mean was set to zero at the first measurement occasion (μ_1) and 0.5 at the second measurement occasion (μ_2). The correlation between the first and second measurement occasions was set at $\rho = 0.3$ and 0.7. The slopes of the items for the latent variable ($\boldsymbol{\beta}$) were generated from a uniform distribution $U(1.1, 2.8)$ [47]. The slopes of the item-specific doublets

(β_s) were specified as 0.5. The thresholds ($\boldsymbol{\alpha}$) for each item were sampled from a uniform distribution $U(-2, 2)$ [48, 49]. The number of response categories (c) for each item was set to five [11].

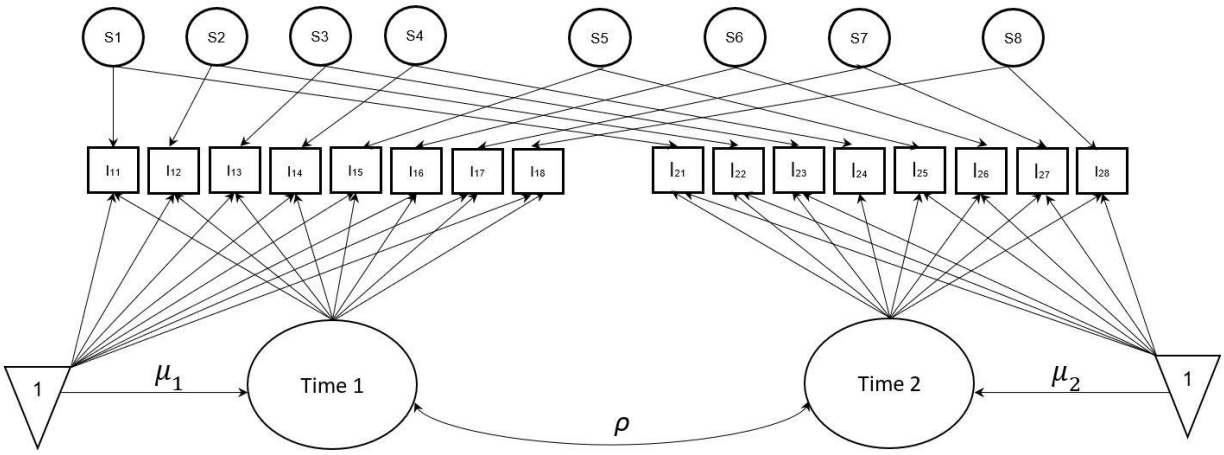


Figure 3-1. The path diagram of a two-tier longitudinal graded response for two measurement occasions, with latent means at the first (μ_1) and second (μ_2) measurement occasions, eight items ($I_{11} - I_{28}$) administered to the same group of individuals, creating two time-specific latent variable for the measurement occasions (Time 1 and Time 2), and eight item-specific doublets ($S_1 - S_8$) capturing the residual correlations.

Item non-response and sample size

Pre-specified amounts (10%, 25% and 50%) of observations were set to have missing responses on the items. We implemented the three different mechanisms of missingness (MCAR, MAR, and MNAR) to remove observations in the simulated dataset [50] using the multivariate amputation method, which generates missing data by assigning a weight to each observation [51]. For example, to generate missing data under the MAR mechanism, the item weights depend only on the observed data; the probability of missingness is calculated assuming a logistic distribution based on the weights. Observations with larger weights have a greater likelihood of missingness, while observations with smaller weights have a lower likelihood of missingness.

This method has been shown to be more appropriate for generating missing data than a stepwise univariate amputation method (i.e., missingness on items are generated one at a time) [52, 53]. The pattern of missingness was specified to be non-monotone to mimic realistic missingness problems. Random samples of size $n = 500$ and 1000 were generated from each of the population models.

Fitted models

Under the first scenario, a fully constrained two-tier longitudinal GRM was fitted to the data prior to estimating the latent variable means. The fitted model was consistent with the invariance condition assumed for the population model. Under the second scenario, two two-tier longitudinal GRMs was fitted to the data. The first fitted model was specified to be consistent with the population model (i.e., the slopes and thresholds of 50% of the items were freely estimated at each measurement occasion). The second fitted model enforced a fully constrained model on data generated under the assumption of partial invariance.

The fitted models in the first and second scenarios were specified to be consistent with the population measurement model. However, in practice since the population measurement is unknown it is crucial to achieve longitudinal measurement invariance to draw valid conclusions about change in latent variables means over time.

Implementation of the methods

All analyses were carried out in *R version 3.5.2.* The two-tier longitudinal single-factor GRM was implemented using the *mirt* package [54]. The NNMF method was implemented using the *NNLM* package [55, 56]. Pre-specified amounts of responses were removed from the items for the three different missing mechanism using the *ampute* function in the *mice* package [40]. The POM method with ten imputations, FIML and NNMF methods were used to address item non-

response. This number of imputations has been shown to be sufficient for up to 50% missing data [25].

Replications

A total of 500 replications were conducted for: (i) each of the 36 simulation conditions obtained by crossing sample size, correlation values for the two time-specific latent variable, pre-specified amounts of missing item responses and missingness mechanisms for the first scenario; and (ii) each of the 36 simulation conditions obtained by crossing sample size, magnitude of inequality, pre-specified amounts of missing item responses and missingness mechanisms for the second scenario.

Performance measures

The performance of the item non-response methods was assessed using absolute percentage bias, mean squared error (MSE) and relative efficiency (RE):

$$APB\left(\hat{\psi}\right) = \frac{\frac{1}{B}\sum_{i=1}^{B}\hat{\psi}^{\,i} - \psi}{\psi} \times 100\%, \tag{7}$$

$$MSE\left(\hat{\psi}\right) = \left(\frac{1}{B}\sum_{i=1}^{B}\hat{\psi}^{\,i} - \psi\right)^{2} + Var\left(\hat{\psi}\right), \tag{8}$$

$$RE_a = \frac{MSE_a\left(\hat{\psi}\right)}{MSE_{NNMF}\left(\hat{\psi}\right)}, \qquad a \in \{\text{CCA, FIML, POM}\}, \tag{9}$$

where ψ is the difference in latent variable means between the first and second measurement occasions, $\hat{\psi}$ is the estimated change in latent variable means between the first and second measurement occasions, and B is the number of simulation replications. Note that $RE_a > 1$ indicates the superiority of NNMF over a, $a \in \{\text{CCA, FIML, POM}\}$. We categorized absolute percentage bias less than 5% as negligible, between 5% and 10% as moderate, and greater than

10% as substantial [12, 57]. We reported the percentage of simulation replications for each condition in each of these categories.

3.4.3 Numeric Example

Real-world data to demonstrate the implementation of the methods were from Winnipeg Regional Health Authority (WRHA) Joint Replacement Registry (JRR) in Manitoba, Canada. This registry contains both general-purpose and condition-specific patient-reported outcomes [58] for patients having total or partial hip or knee arthroplasty surgeries. Patient-reported outcomes data were collected using the 12-item Short-Form (SF-12) survey one-month pre-surgery and one-year post-surgery; we focused on the six mental component summary (MCS) items for patients who had total hip arthroplasty between April 1, 2009, and March 31, 2015, the most recent data available at the time of the analysis. The items of the MCS domain were considered because we anticipated a small effect size that would be sensitive to item non-response [58].

We randomly sampled 1500 patients from the JRR, so that the size of the analytic cohort would be comparable to the largest total sample size in the simulation study. Percentages of non-response were reported for each item. NNMF, FIML, POM and CCA methods were used to address item non-response. The two-tier longitudinal GRM was used to estimate the latent variable means for fully and partially constrained models. The first two items of the MCS domain were constrained for the partially constrained model. We reported the estimated latent variable means, 95% confidence intervals and sample-size adjusted Bayesian information criterion (SABIC) [59, 60] for the models across all the item non-response methods.

3.5 Results

3.5.1 Simulation Study Results

Bias

Scenario 1: Full Invariance. The average absolute percentage bias of the estimated change in latent variable means stratified by time-specific latent variable correlation, and sample size for the CCA, FIML, POM, and NNMF methods are reported in Table 1. The average absolute percentage bias of the estimated change in latent variable means of the fully constrained model when the time-specific latent variable correlation was set to 0.3 ranged between 8.65% and 15.15% for CCA; for FIML the corresponding values was 7.45% and 10.69%; for POM the corresponding values was 7.40% and 10.50%; and for NNMF the corresponding values was 8.20% and 12.25%. When the correlation was set to 0.7, the average estimated change in latent variable means across simulation conditions was approximately 5.0% more than the true change for FIML, POM and NNMF methods.

The magnitude of absolute percentage bias of the estimated change in latent variable means of a fully constrained model stratified by the magnitude of the time-specific latent variable correlation and sample size for the CCA, FIML, POM, and NNMF methods are presented in Table 2. When the sample size was set to 500 and the correlation was 0.3, more than 50.0% of the absolute percentage values for the estimated change in latent variables means were substantial for all the methods under consideration. When the correlation increased to 0.7, less than 20.0% of the absolute percentage bias values for the estimated change in latent variable means were substantial for FIML, POM, and NNMF. More than 50.0% of the absolute percentage bias of the estimated change in latent means were negligible for NNMF when the sample size was set to 1000 and the correlation was 0.7.

Table 3-1. Average absolute percentage bias for the estimated change in latent variable means stratified by the time-specific latent variable correlation and sample size for the first scenario

Sample Size	ρ	Missing Mech.	CCA	FIML	POM	NNMF
500	0.3	MCAR	12.21	10.49	10.25	11.24
		MAR	15.15	10.54	10.36	12.25
		MNAR	12.70	10.69	10.50	12.00
	0.7	MCAR	7.86	5.94	6.04	5.77
		MAR	7.09	5.90	5.98	5.35
		MNAR	7.22	5.81	5.89	5.47
1000	0.3	MCAR	8.65	7.45	7.40	8.20
		MAR	11.66	7.50	7.46	9.19
		MNAR	9.44	7.59	7.56	8.86
	0.7	MCAR	6.31	5.51	5.56	5.07
		MAR	5.38	5.49	5.53	4.56
		MNAR	5.99	5.45	5.50	4.77

MCAR: Missing completely at random; MAR: Missing at random; MNAR: Missing not at random; $\boldsymbol{\rho}$**:** time-specific latent variable correlation; CCA: complete case analysis; FIML: full information maximum likelihood; POM: fully conditional specification proportional odds model; NNMF: non-negative matrix factorization

Table 3-2. Magnitude of absolute percentage bias (%) for the estimated change in latent variable means of the fully constrained model stratified by the time-specific latent variable correlation and sample size for the first scenario

Sample Size	ρ	Missing Mech.	Magnitude of Bias	CCA	FIML	POM	NNMF
500	0.3	MCAR	Negligible	22.9	19.1	19.8	15.7
			Moderate	23.1	29.3	30.0	27.7
			Substantial	54.1	51.6	50.2	56.5
		MAR	Negligible	13.8	18.7	19.2	11.2
			Moderate	18.5	29.1	30.2	25.7
			Substantial	67.7	52.2	50.6	63.1
		MNAR	Negligible	19.4	18.0	19.1	12.3
			Moderate	23.3	29.5	30.3	25.9
			Substantial	57.3	52.5	50.6	61.8
	0.7	MCAR	Negligible	40.3	47.0	46.5	49.1
			Moderate	28.1	35.5	36.2	35.6
			Substantial	31.7	17.5	17.3	15.3
		MAR	Negligible	44.5	48.2	48.1	52.3
			Moderate	31.7	34.7	34.0	36.0
			Substantial	23.8	17.1	17.9	11.7
		MNAR	Negligible	43.0	48.2	48.1	53.2
			Moderate	30.7	35.3	34.7	34.2
			Substantial	26.3	16.5	17.1	12.6
1000	0.3	MCAR	Negligible	32.4	29.2	30.1	24.0
			Moderate	29.7	44.0	43.4	43.5
			Substantial	37.9	26.9	26.5	32.5
		MAR	Negligible	19.0	29.6	29.9	18.5
			Moderate	26.2	43.2	43.8	39.7
			Substantial	54.8	27.2	26.3	41.8
		MNAR	Negligible	28.0	29.3	29.0	20.1
			Moderate	31.4	43.1	43.3	41.5
			Substantial	40.6	27.7	27.7	38.4
	0.7	MCAR	Negligible	46.4	47.6	47.2	54.7
			Moderate	33.5	40.3	40.5	36.8
			Substantial	20.1	12.1	12.3	8.5
		MAR	Negligible	56.3	48.5	48.0	59.5
			Moderate	29.2	39.9	40.1	34.6
			Substantial	14.5	11.7	11.9	5.9
		MNAR	Negligible	49.7	49.4	48.5	59.1
			Moderate	32.0	38.9	39.2	33.6
			Substantial	18.3	11.7	12.3	7.3

MCAR: Missing completely at random; MAR: Missing at random; MNAR: Missing not at random; **ρ:** time-specific latent variable correlation; CCA: complete case analysis; FIML: full information maximum likelihood; POM: fully conditional specification proportional odds model; NNMF: non-negative matrix factorization; negligible bias: absolute percentage bias less than 5%; moderate bias: absolute percentage bias between 5% and 10%; substantial bias: absolute percentage bias greater than 10%

Scenario 2: Partial Invariance. The average absolute percentage bias of the estimated change in latent variable means stratified by magnitude of inequality and sample size for CCA, FIML, POM, and NNMF are presented in Table 3. The average absolute percentage bias of the estimated change in latent variable means for the partially constrained model when the time-specific latent variable correlation was set to 0.7, and sample size was set to 500 ranged between 8.48% and 9.91% for CCA; for FIML the corresponding values were 7.21% and 7.77%; for POM the corresponding values were 7.26% and 7.83%; and for NNMF the corresponding values were 6.74% and 7.48%. When the sample size was increased to 1000, the average absolute percentage bias of the estimated change in latent variable means for CCA ranged between 6.43% and 6.91%; for FIML the corresponding values were 5.38% and 5.59%; for POM the corresponding values were 5.38% and 5.64%; and for NNMF the corresponding values were 4.90% and 5.16%.

The magnitude of the absolute percentage bias of the estimated change in latent variable means of a partially constrained model when the time-specific latent variable correlation was 0.7 are reported in Table 4. When sample size was 500 and magnitude of inequality was set to 0.3, more than one-third of the absolute percentage bias values of the estimated change in latent variable means for CCA were substantial for the three different missing data mechanisms, and less than 30.0% were substantial for FIML, POM, and NNMF. When the sample size was increased to 1000, and magnitude of inequality was set to 0.3, less 50% of the absolute percentage bias values of the estimated change in latent variable means for CCA were negligible for the three different missing data mechanisms, and more than 50% of the absolute percentage bias values were negligible for FIML, POM, and NNMF. When the magnitude of inequality was set to 0.6, similar results were obtained.

Table 3-3. Average absolute percentage bias for the estimated change in latent variable means for $\rho = 0.7$ stratified by magnitude of inequality and sample size for the second scenario

Sample Size	Mag. Of Inequality	Missing Mech.	Partially Constrained Model				Fully Constrained Model			
			CCA	FIML	POM	NNMF	CCA	FIML	POM	NNMF
500	0.3	MCAR	9.21	7.31	7.36	7.05	17.71	17.43	17.20	18.23
		MAR	8.48	7.24	7.33	6.74	19.61	17.36	17.24	19.28
		MNAR	9.21	7.21	7.26	6.80	17.23	17.72	17.56	19.05
	0.6	MCAR	9.60	7.77	7.83	7.48	37.93	37.98	37.70	38.97
		MAR	9.07	7.72	7.77	7.23	37.73	37.84	37.77	40.13
		MNAR	9.91	7.75	7.79	7.37	35.55	38.26	38.10	39.84
1000	0.3	MCAR	6.91	5.59	5.64	5.16	16.98	16.93	16.87	17.78
		MAR	6.43	5.59	5.62	4.93	18.69	16.83	16.87	18.83
		MNAR	6.74	5.53	5.56	4.96	16.53	17.00	17.04	18.48
	0.6	MCAR	6.82	5.43	5.47	5.05	37.94	37.72	37.67	38.75
		MAR	6.67	5.41	5.44	4.90	37.03	37.54	37.68	40.01
		MNAR	6.89	5.38	5.38	4.94	35.34	37.81	37.91	39.66

MCAR: Missing completely at random; MAR: Missing at random; MNAR: Missing not at random; ρ: correlation between the time-specific latent variable; CCA: complete case analysis; FIML: full information maximum likelihood; POM: fully conditional specification proportional odds model; NNMF: non-negative matrix factorization

Table 3-4. Magnitude of absolute percentage bias (%) for the estimated change in latent variable means of a partially constrained model for $\rho = 0.7$ stratified by magnitude of inequality and sample size for the second scenario

Sample Size	Mag. Of Inequality	Missing Mech.	Magnitude of Bias	CCA	FIML	POM	NNMF
500	0.3	MCAR	Negligible	31.9	38.9	38.3	40.7
			Moderate	30.3	33.7	33.6	33.1
			Substantial	37.7	27.4	28.1	26.1
		MAR	Negligible	36.0	39.4	39.0	43.9
			Moderate	30.9	33.5	33.4	32.3
			Substantial	33.1	27.1	27.6	23.8
		MNAR	Negligible	34.7	39.0	38.7	42.9
			Moderate	26.9	34.7	34.4	33.0
			Substantial	38.4	26.3	26.9	24.1
	0.6	MCAR	Negligible	31.0	34.9	36.0	37.6
			Moderate	30.0	34.9	33.5	32.9
			Substantial	39.0	30.1	30.5	29.5
		MAR	Negligible	34.0	36.8	36.7	40.5
			Moderate	27.8	32.6	33.5	31.9
			Substantial	38.2	30.6	29.8	27.5
		MNAR	Negligible	30.5	37.2	37.3	39.2
			Moderate	27.8	32.3	32.1	32.7
			Substantial	41.7	30.5	30.6	28.1
1000	0.3	MCAR	Negligible	46.3	52.0	51.4	56.1
			Moderate	28.6	32.5	33.0	32.0
			Substantial	25.1	15.5	15.6	11.9
		MAR	Negligible	48.9	52.5	52.4	58.4
			Moderate	29.1	32.3	32.0	30.6
			Substantial	22.0	15.2	15.6	11.1
		MNAR	Negligible	45.1	53.8	52.7	58.4
			Moderate	32.6	30.5	32.1	30.4
			Substantial	22.3	15.7	15.2	11.2
	0.6	MCAR	Negligible	46.0	53.6	52.3	57.1
			Moderate	30.1	31.9	33.4	31.3
			Substantial	23.9	14.5	14.3	11.5
		MAR	Negligible	46.4	54.5	53.7	59.2
			Moderate	30.6	31.5	32.0	29.9
			Substantial	23.0	14.1	14.3	10.9
		MNAR	Negligible	44.1	55.1	55.1	59.5
			Moderate	31.3	30.3	30.4	28.9
			Substantial	24.6	14.5	14.5	11.6

MCAR: Missing completely at random; MAR: Missing at random; MNAR: Missing not at random; ρ: correlation between the time-specific latent variable; CCA: complete case analysis; FIML: full information maximum likelihood; POM: fully conditional specification proportional odds model; NNMF: non-negative matrix factorization; negligible bias: absolute percentage bias less than 5%; moderate bias: absolute percentage bias between 5% and 10%; substantial bias: absolute percentage bias greater than 10%

Relative Efficiency

Scenario 1: Full Invariance. Figure 2 shows the RE values for the estimated change in latent variable means for the NNMF method compared with CCA, FIML, and POM methods. The results of the fully constrained model when sample size was 1000 and the time-specific latent variable correlation was $\rho = 0.7$ show an increase in the RE as the percentage of non-response increased. The RE of the NNMF method when compared with the CCA method was greater than 2.0 when 50% of the item responses were MCAR, MAR and MNAR. There was no considerable difference in the RE of the NNMF method compared with the FIML and POM methods when 10% of the item responses were missing under the three missingness mechanisms. The RE of the NNMF method compared with the FIML and POM methods was greater than 1 when 25% or 50% of the item responses were MCAR, MAR, and MNAR. However, this was not the case when correlation between the time-specific latent variable was set to $\rho = 0.3$ (Figure 3).

Scenario 2: Partial Invariance. Figures 4 and 5 show the RE values of the estimated change in latent variable means for the NNMF method, compared with the CCA, FIML, and POM methods for the partially constrained model when magnitude of inequality was 0.3 and 0.6, respectively. The RE when the sample size was set to 1000 and magnitude of inequality was set to 0.3 for the NNMF method when compared with the CCA method was more than 2.0 when 50% of the item responses were MCAR, MAR and MNAR. The RE of the NNMF method compared with the FIML and POM methods was greater than 1 when at least 25% of the item responses were missing under the three missingness mechanisms. Similar results were obtained when the magnitude of inequality was set to 0.6 (Figure 5).

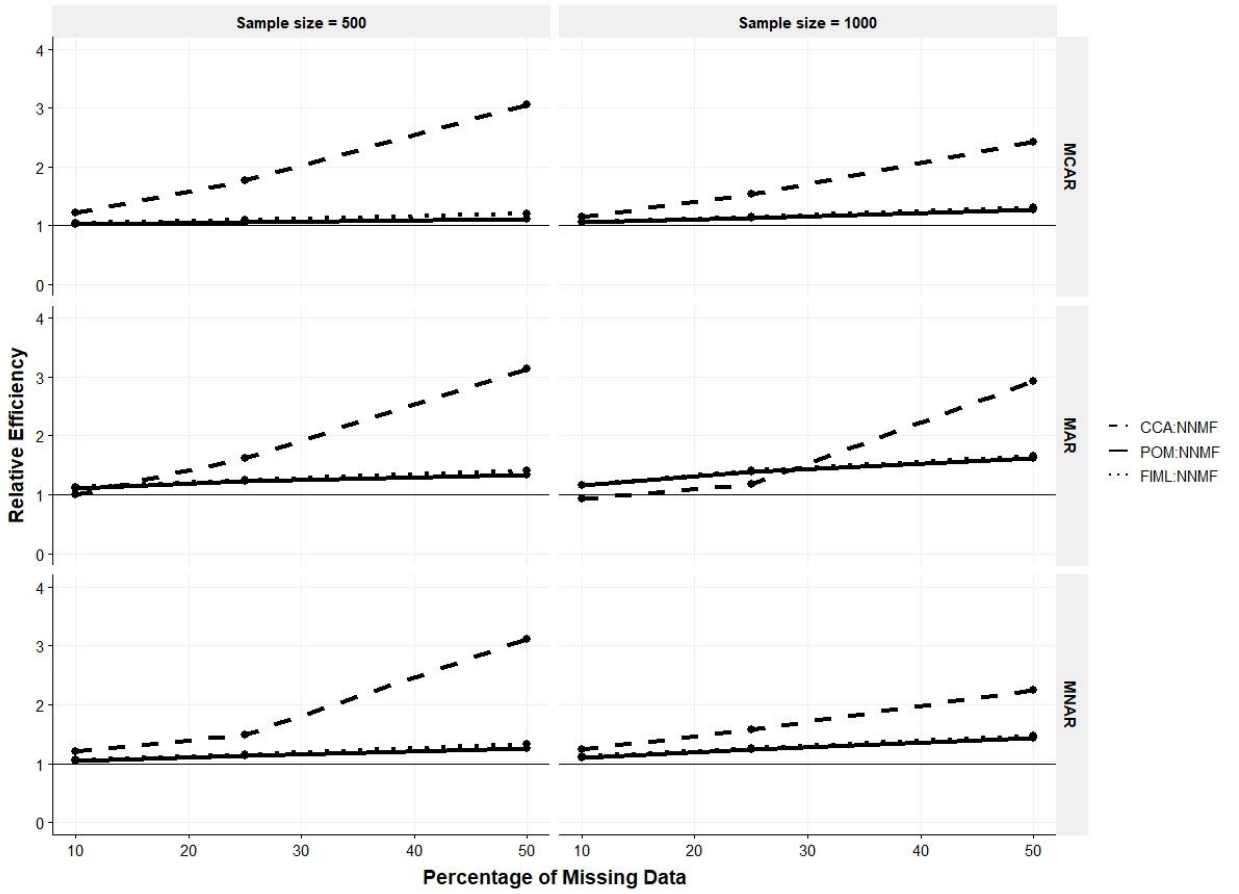


Figure 3-2. Relative efficiency (RE) of the estimated change in latent variable means of a **fully constrained model** for NNMF, compared with CCA, FIML, and POM for three missing data mechanisms, percentage of missingness, sample size, and $\boldsymbol{\rho = 0.7}$ under the **first simulation scenario**. $RE_a > 1$ indicates the superiority of NNMF over a, $a \in \{CCA, FIML, POM\}$. MCAR: Missing completely at random; MAR: Missing at random; MNAR: Missing not at random

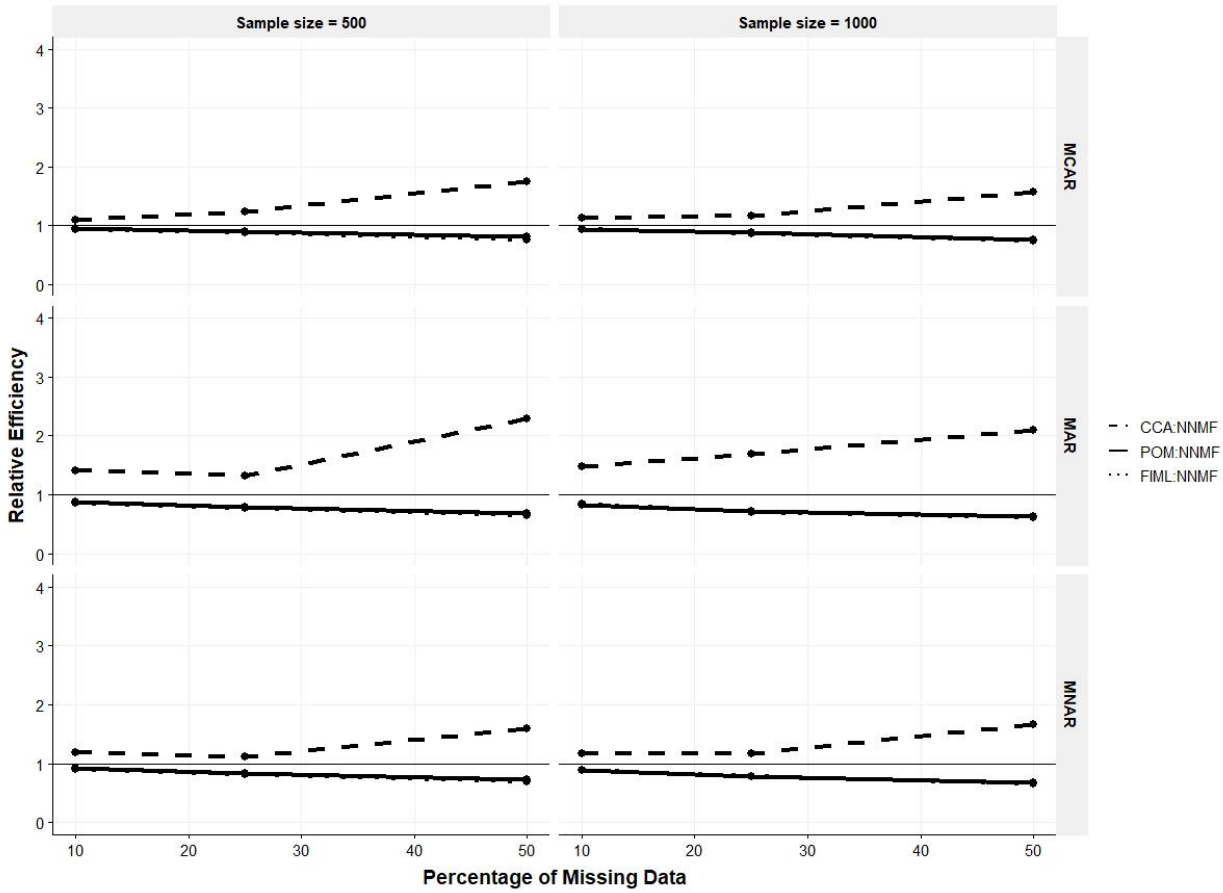


Figure 3-3. Relative efficiency (RE) of the estimated change in latent variable means of a **fully constrained model** for NNMF, compared with CCA, FIML, and POM for three missing data mechanisms, percentage of missingness, sample size, and $\boldsymbol{\rho = 0.3}$ under the **first simulation scenario**. $RE_a > 1$ indicates the superiority of NNMF over a, $a \in \{CCA, FIML, POM\}$. MCAR: Missing completely at random; MAR: Missing at random; MNAR: Missing not at random

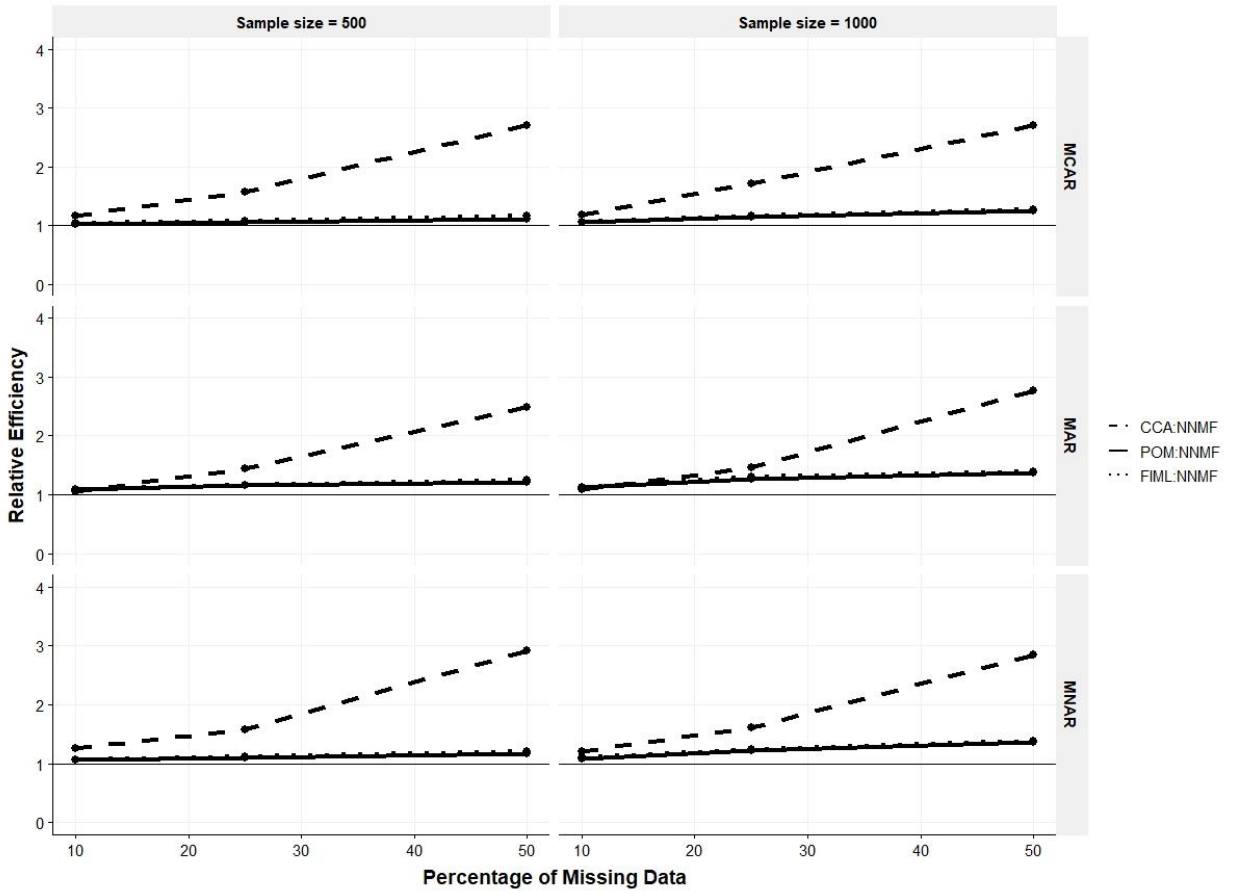


Figure 3-4. Relative efficiency (RE) of the estimated change in latent variable means of a **partially constrained model** for NNMF, compared with CCA, FIML, and POM for three missing data mechanisms, percentage of missingness, sample size, **magnitude of inequality = 0.3** and $\rho = 0.7$ under the **second simulation scenario**. $RE_a > 1$ indicates the superiority of NNMF over a, $a \in \{CCA, FIML, POM\}$. MCAR: Missing completely at random; MAR: Missing at random; MNAR: Missing not at random

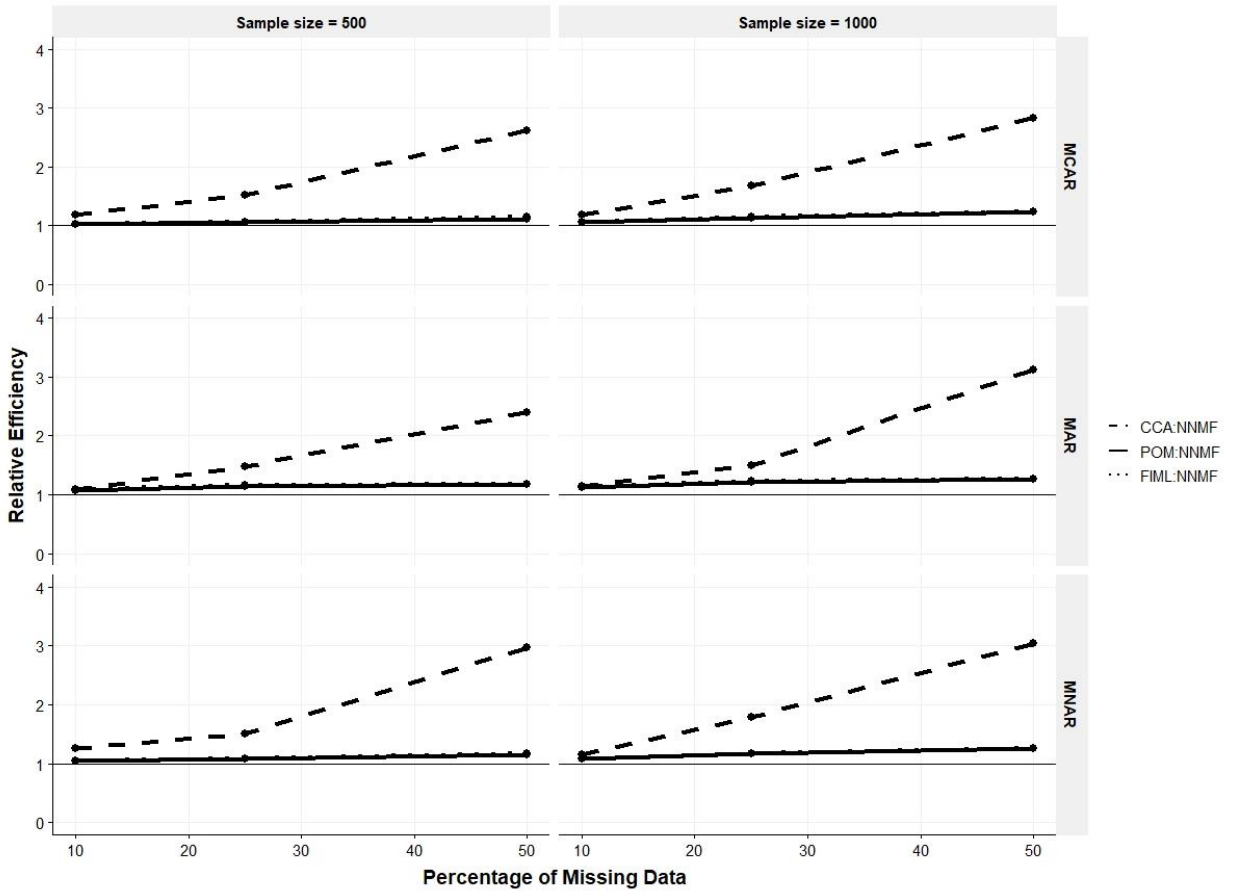


Figure 3-5. Relative efficiency (RE) of the estimated change in latent variable means of a **partially constrained model** for NNMF, compared with CCA, FIML, and POM for three missing data mechanisms, percentage of missingness, sample size, **magnitude of inequality = 0.6** and $\rho = 0.7$ under the **second simulation scenario**. $RE_a > 1$ indicates the superiority of NNMF over a, $a \in \{CCA, FIML, POM\}$. MCAR: Missing completely at random; MAR: Missing at random; MNAR: Missing not at random

3.5.2 Numeric Example

The frequencies and percentages of responses for the MCS items are shown in Table 5. The average non-response rates per item were 16.9% and 31.0% pre- and post-surgery, respectively. The highest percentages of non-response were 19.0% and 31.8% pre- and post-surgery, respectively, for MCS2 (Did work or other activities less carefully than usual). Table 6 shows the estimated change in latent variable means over time with 95% confidence intervals; values of

SABIC are used to describe model fit. The partially constrained model had a better fit (i.e., smaller SABIC) than the fully constrained model for all the methods. The direction of estimated change in the latent MCS mean for the partially constrained model was similar across all the methods. The NNMF method had a smaller estimated change in the mean and the standard error when compared with other item non-response methods.

Table 3-5. Frequencies (%) of item responses on the SF-12 mental component summary (MCS) items for patients in the subsample of the Winnipeg Joint Replacement Registry (*n* = 1500)

Item	**Response options**					
Pre-surgery	**All of the time**	**Most of the time**	**Some of the time**	**A little of the time**	**None of the time**	**No response**
MCS1: Accomplished less	108 (7.2)	217 (14.5)	302 (20.1)	230 (15.3)	378 (25.2)	265 (17.7)
MCS2: Less careful than usual	104 (6.9)	190 (12.7)	269 (17.9)	243 (16.2)	409 (27.3)	285 (19.0)
MCS3: Felt calm and peaceful	64 (4.3)	209 (13.9)	356 (23.7)	551 (36.7)	78 (5.2)	242 (16.1)
MCS4: Have a lot of energy	169 (11.3)	352 (23.5)	430 (28.7)	264 (17.6)	39 (2.6)	246 (16.4)
MCS5: Felt downhearted and depressed	25 (1.7)	97 (6.5)	343 (22.9)	378 (25.2)	411 (27.4)	246 (16.4)
MCS6: Emotional problems	110 (7.3)	230 (15.3)	365 (24.3)	228 (15.2)	334 (22.3)	233 (15.5)
Post-surgery						
MCS1: Accomplished less	21 (1.4)	56 (3.7)	179 (11.9)	201 (13.4)	577 (38.5)	466 (31.1)
MCS2: Less careful than usual	20 (1.3)	47 (3.1)	165 (11.0)	197 (13.1)	594 (39.6)	477 (31.8)
MCS3: Felt calm and peaceful	15 (1.0)	44 (2.9)	165 (11.0)	586 (39.1)	228 (15.2)	462 (30.8)
MCS4: Have a lot of energy	47 (3.1)	145 (9.7)	298 (19.9)	450 (30.0)	94 (6.3)	466 (31.1)
MCS5: Felt downhearted and depressed	25 (1.7)	25 (1.7)	164 (10.9)	286 (19.1)	538 (35.9)	462 (30.8)
MCS6: Emotional problems	19 (1.3)	45 (3.0)	167 (11.1)	183 (12.2)	629 (41.9)	457 (30.5)

Table 3-6. Estimated change in latent variable means over time (with 95% confidence interval) and sample-size adjusted Bayesian information criterion (SABIC)

Model	CCA	FIML	POM	NNMF
Model 1: Fully constrained model				
Means (95% CI)	0.81(0.73, 0.89)	0.96(0.86, 1.05)	0.93(0.85, 1.01)	0.64(0.60, 0.68)
SABIC	19695	30528	40229	37218
Model 2: Partially constrained model				
Means (95% CI)	0.65(0.57, 0.73)	0.61(0.54, 0.68)	0.66(0.59, 0.73)	0.56(0.52, 0.61)
SABIC	19509	30306	39976	36690

CI: Confidence interval; CCA: complete case analysis; FIML: full information maximum likelihood; POM: fully conditional specification proportional odds model; NNMF: non-negative matrix factorization

3.6 Discussion

This study investigated the NNMF method for addressing ordinal item non-response, and compared its performance with the CCA, FIML and POM methods when estimating longitudinal change in latent variable means using simulated data and a real-world numeric example. We examined different scenarios of a longitudinal latent variable measurement model. Across all scenarios, the absolute percentage bias and MSE of the estimated change latent variable means for the CCA method were consistently poor, which corresponds with the findings of previous studies [7, 22, 61]. The absolute percentage bias of the estimated change in latent variable means for all the methods decreased as the time-specific latent variable correlation increased from 0.3 to 0.7, and sample size increased from 500 to 1000.

In the first simulation scenario, where the slopes and thresholds of the items in the population model were constrained to be equal across both measurement occasions, the NNMF method outperformed the FIML and POM methods when the time-specific latent variable correlation was 0.7. The RE of the NNMF method over other methods increased when sample size increased. The best performance of the NNMF method was observed in the large sample size condition (N = 1000) and when the percentage of item non-response was 25% or more. The NNMF method did not perform as well as the FIML and POM methods when the time-specific latent variable correlation was 0.3. This is likely because the NNMF method relies on the dependencies amongst items to predict the missing responses [30, 32, 33]. The performance of the FIML and POM methods was almost the same for all simulation conditions, which is consistent with previous research [23, 24].

In the second simulation scenario, where the slopes and thresholds of 50% of the items in the population model were not constrained equal across measurement occasions, the NNMF method

performed better than the other item non-response methods under consideration when the fitted model was consistent with the population measurement model, regardless of the magnitude of inequality and missingness mechanisms. However, when a fully constrained model was enforced on data generated under the assumption of partial invariance, the performance of the NNMF method was not consistently better than the FIML and POM methods. The RE of the NNMF method when compared with the FIML and POM methods was greater than 1 when 25% or more of the item responses were missing under the different missingness mechanisms.

The NNMF method explores the relationships that exist among items to address item non-response, without making any distributional assumption. However, there are challenges associated with using the NNMF method to address item non-response. The low rank of the response matrix for the NNMF method needs to be specified. We specified it to be one because we considered a single latent variable. A data-driven method, such as the cross-validation method in supervised machine-learning, could be used to select the low rank when there is no prior knowledge about the number of latent variables [55].

3.6.1 Strengths and Limitations

This study has a number of strengths. First, in our simulation we evaluated the performance of the different item non-response methods under fully and partially model parameter constraints scenarios, which was important to understand any potential risk of confounding of our findings by model specification. In addition, we manipulated the dependencies of the latent variable over the two time points, and magnitudes of inequality, which provides a means to compare the performance of the different item non-response methods across a variety of data attributes. Our demonstration and investigation of the NNMF method further strengthen this study because the NNMF method is an unsupervised machine-learning approach and makes no distributional

assumption about the item responses before imputation. Also, the NNMF method is relatively straightforward to implement and it uses the observed responses from all respondents to impute item-level missing data.

There are, however, some limitations to this study. First, like other simulation studies, the simulation conditions we investigated were not exhaustive. Similar to Chen et al. [11] and Sass et al. [62], we limited the number of response categories for each item to five, and the number of items to eight; larger numbers of categories and items would substantially increase computational time for the simulation. Second, we assumed that the fitted models were correctly specified in our simulation. In practice, the parameters of this model need to be tested for invariance using appropriate methods, such as those described by Liu et al. [35]. Finally, the performance of all the item non-response methods considered was studied under a parameter estimation and recovery settings. Inferences about the model specification such as hypothesis testing under the null and alternative would be important to further understand the performance of these methods.

3.6.2 Conclusion

In summary, we compared the performance of item non-response methods when estimating longitudinal change in latent variable means under fully and partial model parameter constraints. Overall, the absolute percentage bias and MSE of all the methods under consideration was satisfactory, except for the CCA method. However, differences in the performance of the item non-response methods were evident. The results reveal better RE for the NNMF method when compared with the FIML and POM methods, especially when sample size was large and the correlation of the latent variable over measurement occasions was strong.

The NNMF method is an alternative method to address item non-response. It offers an advantage over POM where pooling of model fit statistics, such as the chi-square goodness-of-fit statistic to test measurement equivalence are challenging across multiple imputed datasets. Also, it has the advantage of imputing non-response at the item-level; in contrast, the FIML method addresses non-response at the estimation level. However, researchers should be cautious when using the NNMF method in settings where the correlation between the two time-specific latent variable is weak. Future research could investigate adding auxiliary information such as demographic and health status characteristics, or data-driven constraints to the NNMF method in addressing non-response.

Declarations

Ethics approval and consent to participate. This study received ethical approval from the University of Manitoba Health Research Ethics Board.

Consent to participate. Informed written consent was obtained from all participants whose information was used in the analyses of data from the Winnipeg Regional Health Authority Joint Replacement Registry.

Consent for publication. Not applicable. There is no identifying information for participants contained in this manuscript.

Availability of data and material. Not applicable.

Competing interests. The authors declare that they have no competing interests.

Funding. Funding for this study was provided by the Canadian Institutes of Health Research (grant # MOP-142404). OFA is supported by funding from the Visual and Automated Disease Analytics (VADA) Program at the University of Manitoba. LML was supported by a Tier 1 Canada Research Chair in Methods for Electronic Health Data Quality. MJJ acknowledges the Natural Sciences and Engineering Research Council of Canada for partial support. RB acknowledges the research support of the Canadian Institutes of Health Research.

Authors' contributions. All authors conceived the study and prepared the analysis plan. OFA, MJJ and LML conducted the analysis and prepared the draft manuscript. All authors reviewed and approved the final version of the manuscript.

3.7 References

1. Estabrook, R., Cella, D., Zhao, F., Manola, J., DiPaola, R. S., Wagner, L. I., & Haas, N. B. (2018). Longitudinal and dynamic measurement invariance of the FACIT-fatigue scale: an application of the measurement model of derivatives to ECOG-ACRIN study E2805. *Quality of life research*, *27*(6), 1589–1597.

2. Jabrayilov, R., Emons, W. H. M., de Jong, K., & Sijtsma, K. (2017). Longitudinal measurement invariance of the Dutch outcome questionnaire-45 in a clinical sample. *Quality of Life Research*, *26*(6), 1473–1481.

3. de Ayala, R. J. (2009). *The theory and practice of item response theory*. (D. Kenny, Ed.). New York: The Guilford Press.

4. Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologist*. New Jersey: Lawrence Erlbaum Associates, Inc.

5. DeMars, C. (2010). *Item response theory*. New York: Oxford University Press.

6. Baker, F. B., & Kim, S. H. (2004). *Item response theory: parameter estimation techniques*. (F. B. Baker & S.-H. Kim, Eds.) (2nd ed.). Boca Raton: CRC Press.

7. Bell, M. L., & Fairclough, D. L. (2014). Practical and statistical issues in missing data for longitudinal patient-reported outcomes. *Statistical Methods in Medical Research*, *23*(5), 440–459.

8. Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *American Statistical Association*, *90*(431), 1112–1121.

9. Jia, F., & Wu, W. (2019). Evaluating methods for handling missing ordinal data in structural equation modeling. *Behavior Research Methods*, *51*(5), 2337–2355.

10. Wu, W., Jia, F., & Enders, C. (2015). A comparison of imputation strategies for ordinal missing data on Likert scale variables. *Multivariate Behavioral Research*, *50*(5), 484–503.

11. Chen, P. Y., Wu, W., Garnier-Villarreal, M., Kite, B. A., & Jia, F. (2020). Testing measurement invariance with ordinal missing data: a comparison of estimators and missing data techniques. *Multivariate Behavioral Research*, *55*(1), 87–101.

12. Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373.

13. Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (Second Edi.). New York: Wiley.

14. Banks, K. (2015). An introduction to missing data in the context of differential item functioning. *Practical Assessment, Research and Evaluation*, *20*(12), 1–10.

15. Fairclough, A. D. L., & Cella, D. F. (1996). Functional assessment of cancer therapy (FACT-G): Non-response to individual questions. *Quality of Life Research*, *5*(3), 321–329.

16. Eekhout, I., De Vet, H. C. W., Twisk, J. W. R., Brand, J. P. L., De Boer, M. R., & Heymans, M. W. (2014). Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. *Journal of Clinical Epidemiology*, *67*(3), 335–342.

17. Donneau, A. F., Mauer, M., Molenberghs, G., & Albert, A. (2015). A simulation study comparing multiple imputation methods for incomplete longitudinal ordinal data. *Communications in Statistics: Simulation and Computation*, *44*(5), 1311–1338.

18. Sulis, I., & Porcu, M. (2017). Handling missing data in item response theory. assessing the accuracy of a multiple imputation procedure based on latent class analysis. *Journal of Classification*, *34*(2), 327–359.

19. van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, *16*(3), 219–242.

20. Peyre, H., Leplège, A., & Coste, J. (2011). Missing data methods for dealing with missing items in quality of life questionnaires. a comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French 2003 decennial health survey. *Quality of Life Research*, *20*(2), 287–300.

21. Finch, W. H. (2010). Imputation methods for missing categorical questionnaire data: a comparison of approaches. *Journal of Data Science*, *8*(3), 361–378.

22. Ayilara, O. F., Zhang, L., Sajobi, T. T., Sawatzky, R., Bohm, E., & Lix, L. M. (2019). Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. *Health and Quality of Life Outcomes*, *17*(1), 106.

23. Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351.

24. Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, *7*(2), 147–177.

25. Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: a data analyst's perspective. *Multivariate Behavioral Research*, *33*(4), 545–571.

26. Nguyen, C. D., Carlin, J. B., & Lee, K. J. (2017). Model checking in multiple imputation: an overview and case study. *Emerging Themes in Epidemiology*, *14*(1), 1–12.

27. Donneau, A. F., Mauer, M., Lambert, P., Molenberghs, G., & Albert, A. (2015). Simulation-based study comparing multiple imputation methods for non-monotone

missing ordinal data in longitudinal settings. *Journal of Biopharmaceutical Statistics*, *25*(3), 570–601.

28. Taslaman, L., & Nilsson, B. (2012). A framework for regularized non-negative matrix factorization, with application to the analysis of gene expression data. *PLoS ONE*, *7*(11), 1–7.

29. Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In T. K. Leen, T. G. Dietterich, & V. P. Tresp (Eds.), *Advances in Neural Information Processing Systems 13.* (pp. 556–562).

30. Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791.

31. Pauca, V. P., Piper, J., & Plemmons, R. J. (2006). Nonnegative matrix factorization for spectral data analysis. *Linear Algebra and Its Applications*, *416*(1), 29–47.

32. Zhang, S., Wang, W., Ford, J., & Makedon, F. (2006). Learning from incomplete ratings using non-negative matrix factorization. In *Proceedings of the Sixth SIAM International Conference on Data Mining* (pp. 549–553).

33. Mazumder, R., Hastie, T., & Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, *11*, 2287–2322.

34. Wold, H. (1975). Soft modelling by latent variables: the nonlinear iterative partial least squares (NIPALS) approach. *Journal of Applied Probability*, *12*(S1), 117–142.

35. Liu, Y., Millsap, R. E., West, S. G., Tein, J. Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods*, *22*(3), 486–506.

36. Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543.

37. Ayilara, O. F., Sajobi, T. T., Barclay, R., Bohm, E., Jozani Jafari, M., & Lix, L. M. (2022). A comparison of methods to address item non-response when testing for differential item functioning in multidimensional patient-reported outcome measures. *Quality of life Research*, 1–12.

38. Aggarwal, C. C. (2016). *Recommender systems: the textbook*. Switzerland: Springer International Publishing.

39. Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*(434), 473–489.

40. van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67.

41. van Buuren, S. (2012). *Flexible imputation of missing data* (2nd ed.). Boca Raton, FL: CRC Press.

42. Enders, C. K. (2010). *Applied missing data analysis*. *The Guilford Press*. New York.

43. Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75*(4), 581–612.

44. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*(39), 1–38.

45. Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data: An introduction. *Child Development Perspectives*, *4*(1), 5–9.

46. Meade, A. W., Lautenschlager, G. J., & Hecht, J. E. (2009). Establishing measurement equivalence and invariance in longitudinal data with item response theory. *International Journal of Testing*, *5*(3), 279–300.

47. Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in Psychology*, *7*(February).

48. Olsbjerg, M., & Christensen, K. B. (2015). Modeling local dependence in longitudinal IRT models. *Behavior Research Methods*, *47*(4), 1413–1424.

49. De Ayala, R. J. (1994). The influence of multidimensionality on the graded response model. *Applied Psychological Measurement*, *18*(2), 155–170.

50. Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, New Jersey.

51. Schouten, R. M., Lugtig, P., & Vink, G. (2018). Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation*, *88*(15), 2909–2930.

52. Nassiri, V., Molenberghs, G., Verbeke, G., & Barbosa-Breda, J. (2020). Iterative multiple imputation: a framework to determine the number of imputed datasets. *American Statistician*, *74*(2), 125–136.

53. Goretzko, D. (2021). Factor retention in exploratory factor analysis with missing data. *Educational and Psychological Measurement*, 1–21.

54. Chalmers, R. P. (2012). mirt: a multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29.

55. Lin, X. E., & Boutros, P. C. (2020). Optimization and expansion of non-negative matrix factorization. *BMC Bioinformatics*, *21*(1), 1–10.

56. Lin, X. E., & Boutros, P. (2019). NNLM: a package for fast and versatile nonnegative matrix factorization. Retrieved from https://github.com/linxihui/NNLM

57. Flora, D., & Curran, P. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466–491.

58. Zhang, L., Lix, L. M., Ayilara, O., Sawatzky, R., & Bohm, E. R. (2018). The effect of multimorbidity on changes in health-related quality of life following hip and knee arthroplasty. *Bone and Joint Journal*, *100B*(9), 1168–1174.

59. Kutscher, T., Eid, M., & Crayen, C. (2019). Sample size requirements for applying mixed polytomous item response models: results of a Monte Carlo simulation study. *Frontiers in Psychology*, *10*, 2494

60. Lix, L. M., & Ayilara, O. (2022). Latent variable mixture models to address heterogeneity in patient-reported outcome data. *Methods*, *204(August)*, 151-159.

61. Eekhout, I., De Vet, H. C. W., Twisk, J. W. R., Brand, J. P. L., De Boer, M. R., & Heymans, M. W. (2014). Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. *Journal of Clinical Epidemiology*, *67*(3), 335–342.

62. Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: a comparison of estimators. *Structural Equation Modeling*, *21*(2), 167–180.

CHAPTER 4: THE USE OF AUXILIARY VARIABLES IN MISSING DATA METHODS

4.1 Overview

The previous chapters compared the performance of non-negative matrix factorization (NNMF) and several other missing data methods, including full information maximum likelihood (FIML) and multiple imputation (MI) in cross-sectional and longitudinal settings. These methods yield unbiased estimates when data are missing at random (MAR) [1, 2]. However, the assumption that data are MAR cannot be easily tested because the values required for such test are missing.

When data are missing not at random (MNAR), the NNMF, FIML and MI methods produce biased parameter estimates because the distribution of non-response in the data is not considered in these methods [3–5]. Due to the possibility of data being MNAR and the difficulty of testing the MAR assumption in many applications, the NNMF, FIML and MI methods may not be the most reliable approaches in practice. However, these methods can be improved by incorporating variable that is a potential correlate of missingness or a correlate of the missing variable [6]. This variable is called an auxiliary variable. In most cases, an auxiliary variable is ancillary to the substantive research. Integrating an auxiliary variable into the missing data methods can help reduce bias and improve power.

Studies have shown the advantage of integrating an auxiliary variable in the likelihood of several methods including latent growth [7, 8], structural equation models [9] and generalized linear models [10], to improve the accuracy of maximum likelihood analyses. Also, including an auxiliary variable in the imputation model has been used to improve the performance of the MI method in many applications [3, 5, 6, 11, 12]. However, the performance of MI with and without auxiliary variable in patient reported outcome measures (PROMs) research have not been studied.

Chapter 5 investigates the use of auxiliary variable in imputation models and compared the precision and bias with the FIML method, when estimating longitudinal change in PROM scores using true-score model.

Chapter 6 proposes an enhanced weighted NNMF, which uses information from an auxiliary variable to identify clusters in the data and define weights for missing or observed item responses in each cluster before imputation.

4.2 References

1. Lin, X. E., & Boutros, P. C. (2020). Optimization and expansion of non-negative matrix factorization. *BMC Bioinformatics*, *21*(1), 1–10.

2. Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, *7*(2), 147–177.

3. Wang, C., & Hall, C. B. (2010). Correction of bias from non-random missing longitudinal data using auxiliary information. *Statistics in Medicine*, *29*(6), 671–679.

4. Zhang, J., Zhang, Z., & Tao, J. (2021). A Bayesian algorithm based on auxiliary variables for estimating GRM with non-ignorable missing data. *Computational Statistics*, *36*(4), 2643–2669.

5. Mustillo, S., & Kwon, S. (2015). Auxiliary variables in multiple imputation when data are missing not at random. *The Journal of Mathematical Sociology*, *39*(2), 73–91.

6. Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351.

7. Eekhout, I., Enders, C. K., Twisk, J. W. R., de Boer, M. R., de Vet, H. C. W., & Heymans, M. W. (2015). Analyzing incomplete item scores in longitudinal data by including item score information as auxiliary variables. *Structural Equation Modeling*, *22*(4), 588–602.

8. Enders, C. K. (2010). *Applied missing data analysis*. *The Guilford Press*. New York.

9. Enders, C. K. (2008). A note on the use of missing auxiliary variables in full information maximum likelihood-based structural equation models. *Structural Equation Modeling*, *15*(3), 434–448.

10. Ibrahim, J. G., Lipsitz, S. R., & Horton, N. (2001). Using auxiliary data for parameter estimation with non-ignorably missing outcomes. *Applied Statistics*, *50*(3), 361–373.

11. Howard, W. J., Rhemtulla, M., & Little, T. D. (2015). Using principal components as auxiliary variables in missing data estimation. *Multivariate Behavioral Research*, *50*(3), 285–299.

12. Yoo, J. E. (2009). The effect of auxiliary variables and multiple imputation on parameter estimation in confirmatory factor analysis. *Educational and Psychological Measurement*, *69*(6), 929–947.

CHAPTER 5: IMPACT OF MISSING DATA ON BIAS AND PRECISION WHEN ESTIMATING CHANGE IN PATIENT-REPORTED OUTCOMES FROM A CLINICAL REGISTRY

5.1 Overview

This manuscript compares the precision and bias of several missing data methods including complete case analysis, maximum likelihood, multiple imputation with and without an auxiliary variable when estimating change in longitudinal patient reported outcome measures using true-score model. Computer simulation and clinical registry data was used to achieve the study objectives. The results from this study showed that including an auxiliary or supplementary information in the imputation model can increase the precision and reduce bias in the parameter estimates. This paper is a unique addition to the literature on methods for addressing missing data in clinical registries. While previous research has compared different missing data methods, this paper goes further, to investigate the role that supplementary information has to improve the accuracy of the multiple imputation method. Research that focuses on missing data in clinical registries can help understand patients' perspective on improvement in health-related quality of life.

Publication Details

Ayilara, O. F., Zhang, L., Sajobi, T. T., Sawatzky, R., Bohm, E., & Lix, L. M. (2019). Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. *Health and Quality of Life Outcomes*, 17(1), 1-9.

5.2 Abstract

Background: Clinical registries, which capture information about the health and healthcare use of patients with a health condition or treatment, often contain patient-reported outcomes (PROs) that provide insights about the patient's perspectives on their health. Missing data can affect the value of PRO data for healthcare decision-making. We compared the precision and bias of several missing data methods when estimating longitudinal change in PRO scores.

Methods: This research conducted analyses of clinical registry data and simulated data. Registry data were from a population-based regional joint replacement registry for Manitoba, Canada; the study cohort consisted of 5631 patients having total knee arthroplasty between 2009 and 2015. PROs were measured using the 12-item Short Form Survey, version 2 (SF-12v2) at pre- and post-operative occasions. The simulation cohort was a subset of 3000 patients from the study cohort with complete PRO information at both pre- and post-operative occasions. Linear mixed-effects models based on complete case analysis (CCA), maximum likelihood (ML) and multiple imputation (MI) without and with an auxiliary variable (MI-Aux) were used to estimate longitudinal change in PRO scores. In the simulated data, bias, root mean squared error (RMSE), and 95% confidence interval (CI) coverage and width were estimated under varying amounts and types of missing data.

Results: Three thousand two hundred thirty (57.4%) patients in the study cohort had complete data on the SF-12v2 at both occasions. In this cohort, mixed-effects models based on CCA resulted in substantially wider 95% CIs than models based on ML and MI methods. The latter two methods produced similar estimates and 95% CI widths. In the simulation cohort, when 50% of the data were missing, the MI-Aux method, in which a single hypothetical auxiliary variable

was strongly correlated (i.e., 0.8) with the outcome, reduced the 95% CI width by up to 14% and bias and RMSE by up to 50 and 45%, respectively, when compared with the MI method.

Conclusion: Missing data can substantially affect the precision of estimated change in PRO scores from clinical registry data. Inclusion of auxiliary information in MI models can increase precision and reduce bias but identifying the optimal auxiliary variable(s) may be challenging.

5.3 Introduction

Clinical registries are databases that capture information about the health and healthcare use of patients having a specific health condition or healthcare treatment. Patient-reported outcomes (PROs) are increasingly collected in clinical registries because they provide valuable information about the patient's perspectives on their health, including pain, perceived functional abilities, and mental health [1]. PRO data in registries can be a useful tool for clinicians to assess quality of care and improvements in patient health status, beyond what can be captured from objective measures of health status such as complication rates and patient mortality [2]. Clinical registry data have a number of other potential uses, including evaluations of new programs and treatments. Registry data are also used for research. However, clinical registry data collection and evaluation may not always follow the same methods or practices as are used in research studies involving primary data collection [3]. Clinics may also not have the resources needed to routinely and thoroughly check the data for accuracy and completeness.

Studies involving clinical registry data are often longitudinal in nature; for example, they may examine change in PROs before and after an intervention or healthcare treatment [3]. Longitudinal study findings may be strongly influenced by missing data, which can arise when participants die, miss scheduled clinic visits, or fail to respond to clinic questionnaires or interviews. One potential consequence of missing data in a longitudinal study is a loss of power to detect change. Missing data can also result in under- or over-estimation of treatment effects, depending on its characteristics [3–5].

The choice of methods to handle missing data is generally dependent on the missingness mechanism [6–8]. According to Little and Rubin's taxonomy, these mechanisms can be categorized as missing completely at random (MCAR), missing at random (MAR), or missing

not at random (MNAR) [8]. Data are MCAR if the reason for the missingness is unrelated to the outcomes. MAR arises if the reason for dropout depends on the observed outcomes and possibly on observed covariates at any or all occasions before the individual is lost to follow up. The MNAR mechanism depends, in whole or in part, on unobserved measurements.

In longitudinal studies, commonly used missing data methods include list-wise deletion, complete-case analysis (CCA), average available observation carried forward, last observation carried forward, and conditional or unconditional mean imputation [9–11]. However, these methods may result in a loss of statistical power and biased estimates of change, especially when data are MNAR. Other missing data methods, including maximum likelihood (ML) and multiple imputation (MI), which are practical to implement in real-world data and increasingly being adopted, are recommended when the missing data mechanism is ignorable, that is, when the distribution of the missing data indicator is independent of the missing data, conditional on the observed data [11]. Beyond these methods, machine-learning algorithms such as the k-nearest neighbor method, decision trees, and random forest imputation, which are used to construct predictive models to estimate observations that will replace the missing values, can be used when the missing data mechanism is ignorable [12]. However, these machine-learning algorithms may distort the data distribution or introduce spurious associations when not carefully implemented to address missing data [13].

Other types of registries have used ML and MI methods to address missing data. For example, in cancer registries, MI methods based on specific clinical features have been used to impute missing prostate cancer stage information [14]. In a trauma registry, O'Reilly et al. (2012) identified and handled incomplete data using the MI method [15]. In a national weight control registry, Thomas et al. (2014) addressed missing data using the ML method in their evaluation of

the effect of behavior change on weight-loss trajectories [16]. Similarly, in obesity surgery and medical birth registries, missing observations on the outcome variables were addressed using the ML method [17, 18]. However, for all of the studies, the use of these methods is predicated on the assumption that the data are MAR.

When data are MNAR or the missingness mechanism is non-ignorable, methods such as pattern mixture models, shared parameter models, and selection models are recommended [3, 19]. However, these methods are less frequently used because the missing data mechanism must be modeled, and they can be computationally intensive [11].

Another approach to ensure that the assumption about ignorability of the missing data is plausible is to use auxiliary (i.e., supplementary) variables that are potential correlates of missingness and/or the outcome of interest [20]. The use of auxiliary variables related to the outcome of interest may reduce the bias due to missing data in model estimates by adding information associated with missingness to the model. Auxiliary variables are typically found in external data sources. An example of a data source that may contain useful auxiliary variables is administrative health data, which captures information about healthcare use and health status of patients; the advantage of administrative data is that they are routinely collected for purposes of health system management or remuneration, so they are unlikely to have missing values. Auxiliary variables are generally not of direct interest, other than for keeping the assumptions about ignorability of the missing data plausible [20, 21].

Several studies have shown that the theoretical advantage of auxiliary variables is the same for ML and MI methods [3, 20]. However, it is more straightforward to include auxiliary variables without adjusting the analysis model in MI ,when compared to including them in the ML model [21, 22]. Previous studies have compared MI with other missing data methods using simulated

data drawn from a real-world cohort, to preserve the complex relationships amongst the covariates and the outcomes in a longitudinal setting [6, 23]. However, neither the precision nor the bias of MI with and without auxiliary variables in PROs from clinical registry have been previously compared.

The primary purpose of this study was to compare the impact of several missing data methods on the precision of the estimated change in PRO measures in longitudinal data from a clinical registry. A secondary purpose was to use computer simulation to demonstrate the potential effects of including an auxiliary variable in the imputation model on bias and precision of PRO change estimates in longitudinal data.

5.4 Methods

5.4.1 Data Source and Cohorts

Study data were from the population-based Winnipeg Regional Health Authority (WRHA) Joint Replacement Registry. The WRHA is the largest health region in the central Canadian province of Manitoba, which has a population of approximately 1.2 million residents. The Registry captures more than 90% of all hip and knee replacement surgeries performed within the health region and approximately 75% of all replacement surgeries conducted in the province.

Information contained in the Registry includes age, sex, medical conditions, implant details, complications, and both general and condition-specific PROs. These data are collected via self-report and chart abstraction from medical records [24, 25]. Much of the information is collected at a pre-operative assessment conducted approximately one month prior to surgery, with additional data collection at one year following surgery.

The study cohort included all patients in the Registry who had total knee arthroplasty (TKA) between April 1, 2009 and March 31, 2015. Patients with inaccurate data on sex and BMI were excluded. The simulation cohort was comprised of patients from the study cohort who had complete information on PRO measures, demographic, and health status measures.

5.4.2 Study Measures

Generic and condition-specific PROs in the WRHA Joint Replacement Registry are collected via self-report questionnaires completed in the pre-operative assessment clinic and via mailed self-report questionnaires completed one year following surgery. We limited our attention in this study to the generic Short Form Survey version 2 (SF-12v2), a 12-item generic measure of physical and mental well-being. It produces Physical Component Summary (PCS) and Mental Component Summary (MCS) scores, which can range in value from 0 (worst) to 100 (best). Scores are normalized so that values above or below 50 are better or worse, respectively, than their corresponding values in the general population [26].

Demographic information on patient age and sex are extracted from patients' medical records and included in the Joint Replacement Registry. Age was defined at the time of the pre-operative assessment. Body mass index (BMI) was calculated from self-reported weight and height at the pre-operative assessment. Information about comorbid health conditions, such as heart disease, were also obtained via self-report at the pre-operative assessment.

5.4.3 Missing Data Methods

ML and MI methods were selected for use in this study because of their efficient computational requirements and recommendations in the literature for their adoption in practice [11]. The ML method chooses parameter values that assign the maximum possible probability or probability density to the observed data under a well-defined family of parametric probability models. The

probability or probability density of the realized data is the likelihood function [11]. The missing values are removed from the likelihood by a process of summation or integration. These likelihood functions have a complicated form that requires special computational techniques such as expectation maximization [27]. Estimates obtained using this method are unbiased if the missing data are MAR and the statistical model has been correctly specified.

For the MI method, each missing value is replaced by $M > 1$ imputed values. Each value is a Bayesian draw from the conditional distribution of the missing observation given the observed data. The imputations are expected to represent the information about the missing values that is contained in the observed data for the chosen imputation model. MI involves three distinct tasks: (a) missing values are filled in M times to generate M complete data sets, (b) the complete data sets are analyzed using standard procedures, and (c) results from the M analyses are combined into a single inference estimate. The efficiency of the estimate is dependent on the number of imputations and fraction of missing information [28, 29]. Similar to the ML approach, the MI procedure also relies on the assumption that data are MAR. However, the process of handling missing data differs. In MI, missing values are treated in a step that is completely separate from the analysis. This separation has both positive and negative consequences. On the negative side, there is the possibility that a researcher may proceed to analyze the imputed data without considering how the imputations were generated. For example, a model based on a multivariate normal distribution allows pairwise associations among variables but not interactions. Therefore, imputed data set may tend to exhibit interactions that are weaker than those found in the population. On the positive side, the model is flexible and a straightforward process to include auxiliary variables. It is also not necessary to include these auxiliary variables in subsequent

analyses, as their full effect is taken into account during the MI process, and is automatically carried forward into subsequent analyses of the imputed data [20, 30].

5.4.4 Statistical Analysis

Descriptive statistics including means, standard deviations (SD), frequencies, and percentages were used to describe the cohorts at the baseline (i.e., pre-operative) measurement occasion. Patterns of missing data were described for the study cohort using percentages.

A linear mixed-effects model was used to estimate change in SF-12v2 PCS and MCS scores between pre- and post-operative occasions; the choice of models and covariates was based on previous research with these data [31]. Specifically, the model included a random intercept and multiple fixed covariates, including time, age, sex (male [reference], female), and body mass index (BMI < 24.9, 25.0-29.9, 30.0+ [reference], comorbid chronic conditions (including heart disease, depression, high blood pressure, diabetes and back pain (No [reference], Yes)). The two-way interaction of sex and time was included in the model based on preliminary assessments of model fit using penalized likelihood-based fit statistics (e.g., Akaike Information Criterion).

Mixed-effects regression models based on CCA, ML, and MI methods were applied to the study cohort data; separate analyses were conducted for the PCS and MCS. CCA was conducted for the subset of patients who had no missing observations on any variables at either the pre- or post-operative occasions. For the MI method, Markov Chain Monte Carlo (MCMC) sampling of the full predictive distribution was adopted; it assumes a multivariate normal distribution for the imputations. This assumption was descriptively assessed using quantile-quantile plots of the observed values. Ten imputations were conducted, as this number has been shown to be sufficient for achieving a reasonable efficiency for high proportions of missing observations

[29]. All analyses were carried out in R using the *lme* function [32] and multiple imputation by chained equations [33].

5.4.5 Simulation Study

The simulation study was conducted next. In our analysis of the simulation cohort data, we used all variables previously described for the study cohort in addition to a single hypothetical auxiliary variable, *Z*, which was generated from a bivariate normal distribution. Specifically, *Z* was correlated with both the pre- (Y_1) and post-operative (Y_2) scores with ρ = corr (**Y**, *Z*) = 0.2, 0.5 and 0.8, where **Y** = (Y_1, Y_2).

Random samples of size *n* = 1000 were selected from the simulation cohort; mixed-effects models, as specified previously, were applied to PCS scores. Pre-specified amounts (10%, 25% and 50%) of data were removed from the outcome variable via MCAR, MAR, and MNAR mechanisms by modeling the probability of the missing indicator conditional on the outcome variable using a logistic regression model. The ML, MI and MI-Aux (i.e., multiple imputation with *Z* included the imputation model) methods were used to address missingness.

A total of 1000 replications were conducted for each of the 27 simulation conditions, which were obtained by crossing all possible combinations of types and amounts of missingness with the magnitude of correlation of the hypothetical auxiliary variable with the outcome measure. We evaluated bias and error in the regression parameter estimates including the intercept (β_0), which is the estimated average PRO score at the pre-operative occasion, change (β_T) between the pre- and post-operative occasions, and time-sex interaction (β_{TS}). Specifically, we computed standardized bias, root mean squared error (RMSE), 95% confidence interval (CI) coverage, and the average width of the 95% CI for each regression parameter mentioned above. Standardized bias was the ratio of the bias, the difference between the estimates obtained from the model

applied to the random sample with n = 1000 observations, and all data in the simulation cohort, and the SD of the estimates expressed as a percent; smaller values indicate less bias. The RMSE was calculated from the sum of squared bias and variance; smaller values indicate less error. Coverage was calculated as the proportion of the replications for which the 95% CI contained the true value of the parameter of interest; good performance is evident when the actual coverage is approximately equal to the nominal coverage rate of 95%. The average width of the 95% CI was the difference between the upper and lower limits of the interval averaged over the number of replications. Shorter intervals imply greater precision and higher power, provided the 95% CI coverage is high.

5.5 Results

5.5.1 Description of Cohorts and Missing Data

Table 1 describes characteristics of the study and simulation cohorts. The average age of the TKA patients was approximately 67 years in both cohorts. More than half of the patients were obese. The most common chronic conditions were high blood pressure and back pain.

Table 5-1. pre-operative characteristics of the study and simulation cohorts

Characteristic	**Study Cohort (n = 5631)**	**Simulation Cohort (n = 3000)**
Age, mean (SD)	66.7 (9.9)	67.0 (9.2)
Sex (female)	3307 (58.7)	1754 (58.4)
Body Mass Index < 24.9 (underweight or normal weight)	632 (11.2)	314 (10.5)
25.0 - 29.9 (overweight)	1696 (30.1)	912 (30.4)
> 30 (obese)	3305 (58.7)	1774 (59.1)
High Blood Pressure	3020 (53.5)	1613 (53.8)
Back Pain	2074 (36.8)	1080 (36.0)
Diabetes	992 (17.6)	494 (16.5)
Depression	761 (13.5)	322 (10.7)
Heart Disease	663 (11.8)	329 (11.0)

Frequencies and percentages [n (%)] reported for all categorical variables; SD – Standard Deviation.

The patterns of missing data in the study cohort is reported in Table 2. Overall, just 57.4% of the cohort had complete data at both pre- and post-operative occasions. Almost one-third of this cohort had missing data at the post-operative occasion only.

Table 5-2. Missing data patterns in the study cohort

Pattern	Pre-Operative	Post-Operative	%
1	O	O	57.4
2	O	X	32.3
3	X	O	5.9
4	X	X	4.4

O – Observed; X – Missing.

5.5.2 Regression Results for the Study Cohort

The mixed-effects regression model results for the study cohort on both the SF-12v2 PCS and MCS measures for the intercept, time, and time-sex effects are presented in Table 3. Parameter estimates, standard errors and 95% CI width are provided for the CCA, ML and MI methods. Overall, the three methods did not differ on statistical significance of the parameter estimates for the intercept, time, and time-sex. However, there were differences amongst the methods for the coefficients of age, diabetes, and heart disease in the model for MCS, which were not statistically significant for the CCA method but were statistically significant for the ML and MI methods (estimates not shown). Overall, the CCA method yielded 95% CIs that were substantially wider than for the ML and MI methods. ML and MI produced similar estimates and 95% CI widths (see Table 3).

Table 5-3. Mixed-effect regression model parameter estimates for the SF-12v2 PCS and MCS scores

		CCA		ML		MI	
	Model Effect	**Estimate (SE)**	**CI Width**	**Estimate (SE)**	**CI Width**	**Estimate (SE)**	**CI Width**
PCS	β_0	**37.85 (1.10)**	4.32	**35.91 (0.86)**	3.36	**37.72 (0.92)**	3.65
	β_T	**12.42 (0.30)**	1.18	**12.67 (0.27)**	1.04	**12.57 (0.26)**	1.03
	β_{TS}	**-1.03 (0.39)**	1.53	**-1.35 (0.35)**	1.36	**-1.34 (0.35)**	1.39
MCS	β_0	**54.20 (1.32)**	5.18	**50.86 (1.07)**	4.19	**51.15 (1.07)**	4.22
	β_T	-0.24 (0.29)	1.13	0.01 (0.27)	1.06	0.004 (0.25)	0.99
	β_{TS}	**1.13 (0.38)**	1.48	**1.34 (0.35)**	1.39	**1.48 (0.33)**	1.32

CCA – Complete Case Analysis; ML – Maximum Likelihood; MI – Multiple Imputation; HBP – High Blood Pressure; CI Width - Width of the 95% confidence interval; SE – Standard Error; Boldface font indicates statistically significant estimates at $\alpha = 0.05$.

5.5.3 Simulation Study Results

The performance measures for the computer simulation, including the standardized bias, RMSE, and average width of the 95% CI for the CCA, ML, and MI methods are reported in Table 4. The 95% CI coverage (not reported) for the CCA, ML and MI methods ranged between 95% and 97% when data were MCAR, between 91% and 97% when data were MAR and between 60% and 93% when data were MNAR. The lowest number in each of these sets of values corresponds to the case when 50% of the data were missing.

The RMSE and the average width of the 95% CI increased as the rate of missingness increased, reflecting the expected loss of information that occurs with increased rates of missing data. Under the different missing data mechanisms, the average 95% CI width and RMSE obtained from the ML and MI methods were similar for all main effects, but not for the two-way interaction. When 25% to 50% of the data were missing, the average width of the 95% CI from the ML method was marginally narrower than the width for the MI method. The standardized bias for the CCA, ML and MI methods when data were MNAR was twice the size of the bias observed when the data were missing because of MCAR and MAR mechanisms. As the rate of

missingness increased, the standardized bias also increased. However, when the missing data were MCAR, the standardized bias for the MI and ML methods were larger than for the CCA method, while the RMSE of the CCA method was substantially larger than for the MI and ML methods.

Simulation results for the MI and MI-Aux methods are reported in Table 5. Including the hypothetical auxiliary variable in the imputation model reduced the average width of the 95% CI as its correlation with the outcome variable increased. The size of the reduction increased as the rate of missing data increased. When 50% of the data was missing, we obtained a reduction of up to 14% and 4% in the average 95% CI width by including the hypothetical auxiliary variable with $\rho = 0.8$ and 0.5, respectively. However, there was no significant reduction in the average 95% CI width when the hypothetical auxiliary variable with $\rho = 0.2$ was included in the model. Similarly, a reduction of up to 5% and 2% in the average 95% CI width was observed when the percentage of missing data was 25% and 10% respectively.

Inclusion of the hypothetical auxiliary variable in the imputation model reduced the bias and RMSE, particularly in cases where the rate of missing data was high and $\rho = 0.8$. When 50% of the data was missing via an MNAR mechanism, the bias reduction was over 50% and RMSE decreased by up to 45%.

Table 5-4. Simulation performance measures for complete case analysis (CCA), maximum likelihood (ML) and multiple imputation (MI)

Missing Mechanism	Model Parameter	Performance Measure	10% Missing			25% Missing			50% Missing		
			CCA	**ML**	**MI**	**CCA**	**ML**	**MI**	**CCA**	**ML**	**MI**
MCAR	β_0	Bias	5.57	6.90	7.53	2.03	1.09	3.97	-0.19	1.51	6.23
		RMSE	1.89	1.83	1.84	2.14	1.96	1.97	2.68	2.19	2.23
		CI Width	8.28	8.09	8.12	9.06	8.48	8.56	11.14	9.31	9.50
	β_T	Bias	-2.91	-3.29	-11.63	-3.47	-1.79	-20.61	-1.90	-2.21	-39.96
		RMSE	0.47	0.47	0.45	0.54	0.52	0.49	0.68	0.60	0.53
		CI Width	2.18	2.15	2.15	2.38	2.29	2.31	2.92	2.62	2.64
	β_{TS}	Bias	3.21	3.27	15.30	3.30	0.89	29.48	-4.47	-2.62	60.89
		RMSE	0.59	0.59	0.55	0.68	0.66	0.56	0.86	0.76	0.58
		CI Width	2.85	2.81	2.81	3.12	3.00	3.00	3.81	3.43	3.32
MAR	β_0	Bias	14.08	6.19	7.17	11.42	0.56	2.01	-13.98	-9.28	-6.69
		RMSE	1.87	1.81	1.81	2.08	1.91	1.92	2.59	2.12	2.16
		CI Width	8.21	8.05	8.08	8.89	8.39	8.47	10.67	9.12	9.24
	β_T	Bias	20.71	14.92	9.90	42.60	32.55	21.65	86.01	64.71	43.61
		RMSE	0.47	0.46	0.44	0.56	0.52	0.47	0.81	0.66	0.52
		CI Width	2.14	2.12	2.13	2.30	2.23	2.25	2.72	2.51	2.56
	β_{TS}	Bias	11.25	10.53	21.15	21.13	18.10	45.06	10.25	3.52	64.01
		RMSE	0.60	0.59	0.56	0.69	0.66	0.59	0.85	0.75	0.59
		CI Width	2.83	2.80	2.81	3.09	2.98	2.98	3.79	3.41	3.31
MNAR	β_0	Bias	-52.67	-41.61	-39.78	-111.56	-82.51	-80.01	-171.53	-105.67	-102.79
		RMSE	2.05	1.89	1.89	3.08	2.45	2.43	5.23	3.16	3.17
		CI Width	8.13	7.95	7.97	8.79	8.27	8.34	10.70	9.16	9.35
	β_T	Bias	-1.46	7.61	-2.71	-25.40	-9.78	-36.13	-73.72	-47.93	-110.44
		RMSE	0.47	0.47	0.45	0.57	0.53	0.51	0.87	0.69	0.76
		CI Width	2.19	2.15	2.17	2.42	2.32	2.34	3.04	2.70	2.73
	β_{TS}	Bias	-17.53	-18.53	-5.93	-32.42	-34.40	-3.92	-45.61	-45.24	17.64
		RMSE	0.59	0.59	0.54	0.72	0.69	0.54	0.99	0.88	0.53
		CI Width	2.85	2.81	2.82	3.13	3.01	3.02	3.90	3.50	3.41

CI Width - Average width of the 95% confidence interval; Bias – Standardized bias; RMSE – Root mean squared error.

Table 5-5. Simulation performance measures for multiple imputation (MI) without and with an auxiliary variable (MI-Aux)

Missing Mechanism	Model Parameter	Performance Measure	10% Missing				25% Missing				50% Missing			
			MI	MI-Aux			MI	MI-Aux			MI	MI-Aux		
				ρ = 0.2	ρ = 0.5	ρ = 0.8		ρ = 0.2	ρ = 0.5	ρ = 0.8		ρ = 0.2	ρ = 0.5	ρ = 0.8
MCAR	β_0	Bias	4.48	3.75	3.70	3.29	2.31	3.05	3.25	0.74	3.89	4.81	5.05	4.36
		RMSE	1.78	1.78	1.77	1.74	1.88	1.89	1.85	1.77	2.19	2.18	2.11	1.94
		CI Width	8.11	8.09	8.03	7.92	8.58	8.53	8.39	8.09	9.50	9.42	9.10	8.46
	β_T	Bias	-2.16	-2.06	-0.54	2.18	-19.72	-21.47	-21.49	-12.81	-36.70	-37.14	-31.22	-16.90
		RMSE	0.46	0.46	0.46	0.45	0.47	0.46	0.46	0.44	0.52	0.53	0.52	0.47
		CI Width	2.16	2.16	2.15	2.12	2.32	2.31	2.29	2.21	2.65	2.64	2.57	2.38
	β_{TS}	Bias	4.18	3.84	2.61	-1.39	29.38	32.62	33.04	21.74	61.45	62.28	52.82	29.08
		RMSE	0.56	0.56	0.56	0.56	0.57	0.57	0.57	0.56	0.60	0.60	0.60	0.58
		CI Width	2.82	2.82	2.81	2.77	3.00	3.00	2.96	2.88	3.34	3.31	3.25	3.05
MAR	β_0	Bias	-1.29	-0.56	0.67	3.06	-11.52	-10.87	-10.59	-8.20	-26.62	-19.67	-10.71	4.93
		RMSE	1.77	1.76	1.76	1.74	1.87	1.88	1.84	1.76	2.26	2.24	2.13	1.95
		CI Width	8.05	8.04	8.01	7.93	8.44	8.45	8.32	8.10	9.34	9.31	9.03	8.44
	β_T	Bias	9.13	6.53	4.63	3.17	-4.55	-11.15	-18.72	-18.78	-14.41	-14.83	-10.96	-2.48
		RMSE	0.46	0.46	0.46	0.46	0.45	0.46	0.46	0.45	0.52	0.51	0.50	0.47
		CI Width	2.15	2.15	2.14	2.12	2.31	2.30	2.28	2.21	2.62	2.60	2.54	2.35
	β_{TS}	Bias	-8.19	-6.72	-6.81	-5.64	15.21	20.60	24.92	21.05	31.05	30.19	22.37	6.38
		RMSE	0.57	0.57	0.57	0.57	0.54	0.55	0.56	0.56	0.55	0.54	0.55	0.57
		CI Width	2.83	2.82	2.81	2.77	3.02	3.02	2.99	2.89	3.34	3.32	3.25	3.05
MNAR	β_0	Bias	-42.75	-40.05	-31.76	-13.71	-86.54	-84.55	-73.35	-42.67	-102.96	-99.42	-89.10	-53.11
		RMSE	1.89	1.87	1.81	1.73	2.49	2.44	2.26	1.90	3.17	3.10	2.80	2.20
		CI Width	7.97	7.97	7.90	7.84	8.33	8.32	8.20	7.97	9.37	9.27	8.95	8.35
	β_T	Bias	7.95	6.41	3.46	1.73	-31.14	-34.74	-35.36	-25.04	-101.73	-99.73	-80.58	-41.96
		RMSE	0.45	0.45	0.45	0.45	0.48	0.48	0.48	0.45	0.74	0.73	0.65	0.51
		CI Width	2.17	2.16	2.16	2.12	2.34	2.35	2.31	2.22	2.74	2.71	2.63	2.40
	β_{TS}	Bias	-16.91	-14.34	-8.65	-3.75	-5.30	1.39	11.72	16.71	16.66	18.66	15.27	8.17
		RMSE	0.55	0.55	0.55	0.56	0.53	0.54	0.55	0.55	0.54	0.54	0.55	0.57
		CI Width	2.82	2.82	2.80	2.77	3.01	3.02	2.98	2.88	3.41	3.39	3.30	3.07

CI Width – Average width of the 95% confidence interval; Bias – Standardized bias; RMSE – Root mean squared error.

5.6 Discussion

This study used a real-world numeric example and computer simulation to compare several methods for missing data when estimating change in PRO scores from a joint replacement clinical registry. In the numeric example, we investigated the effect of missing data methods on the precision of estimates of change in pre- and post-operative PROs. Standard errors were consistently larger for the CCA method when compared with ML and MI methods. The ML and MI methods produced consistent parameter estimates and standard errors. This is because the imputation and analysis models are similar for both methods [11, 20].

The simulation study investigated the potential benefit of using a supplementary variable on the bias and precision of the MI method. The simulation focused on the effects of a single hypothetical auxiliary variable, although in practice there may be more than one auxiliary variable included in the MI model [20]. The impact on bias and precision was substantial when the amount of missing data was large, and when the correlation between the hypothetical auxiliary variable and the outcome of interest was high. When missingness on the outcome of interest was ignorable, inclusion of an auxiliary variable that was strongly associated with our outcome variable added extra information to the imputation model, which is in agreement with the recommendation of the International Society of Arthroplasty Registries on how to deal with missing data in arthroplasty registries [34]. As a result of this inclusion, we obtained a significant reduction in standard errors, and consequently increased the precision of our analysis, which is consistent with previous research [3, 20, 22]. Furthermore, including an auxiliary variable in the imputation model helped moderate the amount of bias and size of the RMSE when missingness was non-ignorable. Thus, our results are comparable to what would be expected under an ignorable missingness mechanism [20].

Clinical registries may include variables that are correlated with missingness and could therefore be included as auxiliary variables in MI models. However, many variables in clinical registries will have the same pattern of missingness as the outcome of interest. This was true for the clinical registry data used in the current study. Thus, in order to adopt a MI approach with auxiliary variables, the researcher should link the registry data to another data source that has complete information on variables thought to be associated with missingness. For example, it may be possible to link clinical registry data with administrative health data containing measures of healthcare use or diagnoses for comorbid health conditions that may be associated with the presence of missing observations [35, 36]. As well, these data may also contain information about physician characteristics, which may also influence missing data for patients captured in clinical registries. For example, some physicians may focus on more or less complex patients; comorbid characteristics, which are likely to be more common in more complex patients, may have a strong impact on patient dropout/loss to follow-up.

This study has a number of strengths. First, we used a combination of computer simulation and a real-world numeric example to examine the effect of the missing data method on estimates of change in PRO scores. The use of simulated data drawn from the study cohort ensures that the complex relationships amongst the covariates and the outcomes are preserved, which facilitates understanding of the impacts of missing data in real-world settings [23]. We examined change in both the SF-12v2 PCS and MCS scores in our numeric example, and PCS only in our simulation study as we expect the performance measures to have similar patterns across outcomes. There are, however, some limitations to this study. In our simulation study, we considered only the hypothetical situation of using a single auxiliary variable in the imputation model due to the substantial computation time for the simulation study. Also, we only considered the case where

the relationship between the auxiliary variable and outcome of interest was linear. It is possible to have a scenario where the relationship is non-linear. Moreover, we did not include an auxiliary variable in our numeric example. The linkage of the WRHA Joint Replacement Registry to sources of auxiliary variables, such as administrative health data, requires data access approvals and a health information privacy impact assessment. Moreover, the choice of one or more potential auxiliary measures to include in a MI model is not a straightforward process; both theoretical and practical considerations must be addressed, which is beyond the scope of the current paper [20].

5.6.1 Conclusion

In summary, we examined the impact of different missing data mechanisms and an auxiliary variable on the bias and precision in estimating change over time in PROs. Our simulation results showed that using auxiliary information in the imputation model can increase the precision and reduce the bias of parameter estimates, especially in cases where the percentage of missing data is high.

In the absence of an auxiliary variable, the simulation results revealed that the ML method is more precise in estimating longitudinal change in PRO measures than the MI method, especially when there is complete data on the covariates. However, MI offers an advantage of straightforward inclusion of one or more auxiliary variables in the imputation model over the ML method. Under the expectation of inevitable missing data when conducting a longitudinal study, complete auxiliary information should be collected, such as other measures of the PRO of interest and/or variables that may be associated with the outcome. Our results showed a consistent pattern in all the scenarios considered. Therefore, we recommend that in the presence

of missing data, initial analyses should be conducted assuming MAR and then sensitivity analyses should be conducted assuming MNAR.

List of Abbreviations

PRO: Patient Reported Outcome

MCAR: Missing Completely at Random

MAR: Missing at Random

MNAR: Missing not at Random

ML: Maximum Likelihood

MI: Multiple Imputation

MI-Aux: Multiple Imputation with Auxiliary Variable

WRHA: Winnipeg Regional Health Authority

TKA: Total Knee Arthroplasty

SF-12v2: Short Form Survey version 2

PCS: Physical Component Summary

MCS: Mental Component Summary

CCA: Complete Case Analysis

MCMC: Markov Chain Monte Carlo

RMSE: Root Mean Squared Error

CI: Confidence Interval

Declarations

Competing interests. None declared

Availability of data and material. Data used in this article were derived from health data as a secondary source. The data were provided under specific data sharing agreements only for the approved use. The original source data are not owned by the researchers and as such cannot be provided to a public repository. The original data source and approval for use has been noted in the acknowledgments of the article. Where necessary and with appropriate approvals, source data specific to this article or project may be reviewed with the consent of the original data providers, along with the required privacy and ethical review bodies.

Authors' contributions. All authors conceived the study and prepared the analysis plan. OFA, and LML conducted the analysis and prepared the draft manuscript. All authors reviewed and approved the final version of the manuscript.

Ethics approval. This study received ethical approval from the University of Manitoba Health Research Ethics Board.

Consent to participate. Informed written consent was obtained from all participants whose information was used in the analyses of data from the Winnipeg Regional Health Authority Joint Replacement Registry.

5.7 References

1. Franklin, P. D., Ayers, D. C., & Berliner, E. (2018). The essential role of patient-centered registries in an era of electronic health records. *NEJM Catalyst*.

2. Johnston, B. C., Patrick, D. L., Thorlund, K., Busse, J. W., da Costa, B. R., Schünemann, H. J., & Guyatt, G. H. (2013). Patient-reported outcomes in meta-analyses - part 2: methods for improving interpretability for decision-makers. *Health and Quality of Life Outcomes*, *11*(211), 1–9.

3. Bell, M. L., & Fairclough, D. L. (2014). Practical and statistical issues in missing data for longitudinal patient-reported outcomes. *Statistical Methods in Medical Research*, *23*(5), 440–459.

4. Schafer, J. L. (1997). *Analysis of incomplete multivariate data* (1st ed.). London, United Kingdom: Chapman & Hall Ltd.

5. Molenberghs, G., & Kenward, M. G. (2007). *Missing data in clinical studies*. West Sussex: John Wiley & Sons.

6. Peyre, H., Leplège, A., & Coste, J. (2011). Missing data methods for dealing with missing items in quality of life questionnaires. A comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French 2003 decennial health survey. *Quality of Life Research*, *20*(2), 287–300.

7. Myers, W. R. (2000). Handling missing data in clinical trials: an overview. *Drug Information Journal*, *34*, 525–533.

8. Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

9. Gomes, M., Gutacker, N., Bojke, C., & Street, A. (2016). Addressing missing data in patient-reported outcome measures (PROMS): implications for the use of PROMS for comparing provider performance. *Health Economics*, *25*(5), 515–528.

10. White, I. R., & Carlin, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*, *29*(28), 2920–2931.

11. Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, *7*(2), 147–177.

12. Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine*, *50*(2), 105–115.

13. Beretta, L., & Santaniello, A. (2016). Nearest neighbor imputation algorithms: a critical evaluation. *BMC Medical Informatics and Decision Making*, *16*(Suppl 3), 197–208.

14. Parry, M. G., Sujenthiran, A., Cowling, T. E., Charman, S., Nossiter, J., Aggarwal, A., … van der Meulen, J. (2019). Imputation of missing prostate cancer stage in English cancer registry data based on clinical assumptions. *Cancer Epidemiology*, *58*, 44–51.

15. O'Reilly, G. M., Cameron, P. A., & Jolley, D. J. (2012). Which patients have missing data? An analysis of missingness in a trauma registry. *Injury*, *43*(11), 1917–1923.

16. Thomas, J. G., Bond, D. S., Phelan, S., Hill, J. O., & Wing, R. R. (2014). Weight-loss maintenance for 10 years in the national weight control registry. *American Journal of Preventive Medicine*, *46*(1), 17–23.

17. Dreber, H., Thorell, A., Torgerson, J., Reynisdottir, S., & Hemmingsson, E. (2018). Weight loss, adverse events and loss-to follow- up after gastric bypass in young versus older adults: a Scandinavian obesity surgery registry study. *Surgery for Obesity and Related Disease*, *14*(9), 1319–1326.

18. Lenters, V., Iszatt, N., Forns, J., Ko, A., & Legler, J. (2019). Early-life exposure to persistent organic pollutants (OCPs, PBDEs, PCBs, PFASs) and attention-deficit / hyperactivity disorder: a multi-pollutant analysis of a Norwegian birth cohort. *Environment International*, *125*, 33–42.

19. Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, *88*(421), 125–134.

20. Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351.

21. Eekhout, I., Enders, C. K., Twisk, J. W. R., de Boer, M. R., de Vet, H. C. W., & Heymans, M. W. (2015). Analyzing incomplete item scores in longitudinal data by including item score information as auxiliary variables. *Structural Equation Modeling*, *22*(4), 588–602.

22. Wang, C., & Hall, C. B. (2010). Correction of bias from non-random missing longitudinal data using auxiliary information. *Statistics in Medicine*, *29*(6), 671–679.

23. Kalaycioglu, O., Copas, A., King, M., & Omar, R. Z. (2016). A comparison of multiple-imputation methods for handling missing data in repeated measurements observational studies. *Journal of the Royal Statistical Society*, *179*(3), 683–706.

24. Singh, J., Politis, A., Loucks, L., Hedden, D. R., & Bohm, E. R. (2016). Trends in revision hip and knee arthroplasty observations after implementation of a regional joint replacement registry. *Canadian Journal of Surgery*, *59*(5), 304–310.

25. Rolfson, O., Rothwell, A., Sedrakyan, A., Chenok, K. E., Bohm, E., Bozic, K. J., & Garellick, G. (2011). Use of patient-reported outcomes in the context of different levels of data. *The Journal of Bone and Joint Surgery*, *93 Suppl 3*, 66–71.

26. Ware, J. E., Kosinski, M., & Keller, S. D. (1996). A 12-Item short-form health survey : construction of scales and preliminary tests of reliability and validity. *Medical Care*, *34*(3), 220–233.

27. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*(39), 1–38.

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

29. Raghunathan, T. E., Lepkowski, J. M., & van Hoewyk, J. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27*(1), 85–95.

30. Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: a data analyst's perspective. *Multivariate Behavioral Research*, *33*(4), 545–571.

31. Zhang, L., Lix, L. M., Ayilara, O., Sawatzky, R., & Bohm, E. R. (2018). The effect of multimorbidity on changes in health-related quality of life following hip and knee arthroplasty. *Bone and Joint Journal*, *100B*(9), 1168–1174.

32. Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2018). nlme: Linear and nonlinear mixed effects model.

33. van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67.

34. Rolfson, O., Bohm, E., Franklin, P. D., Lyman, S., Denissen, G., Dawson, J., … Dunbar, M. (2016). Patient-reported outcome measures in arthroplasty registries. *Acta Orthopaedica*, *87*(Suppl 1), 9–23.

35. Norris, C. M., Ghali, W. A., Knudtson, M. L., Naylor, C. D., & Saunders, L. D. (2000). Dealing with missing data in observational health care outcome analyses. *Journal of Clinical Epidemiology*, *53*, 377–383.

36. Southern, D. A., Norris, C. M., Quan, H., Shrive, F. M., Gallbraith, D. P., Humphries, K., … Ghali, W. A. (2008). An administrative data merging solution for dealing with missing data in a clinical registry: adaptation from ICD-9 to ICD-10. *BMC Medical Research Methodology*, *8*(1), 1–9.

CHAPTER 6: ENHANCED WEIGHTED NON-NEGATIVE MATRIX FACTORIZATION APPROACH FOR ADDRESSING ITEM NON-RESPONSE IN PATIENT-REPORTED OUTCOME MEASURES

6.1 Overview

This manuscript proposes an enhanced weighted non-negative matrix factorization, which uses the observed item responses as auxiliary or supplementary variable to define weights before item level imputation. The Type I error rate and statistical power of the proposed method was compared with several ordinal item non-response methods including listwise deletion, full information maximum likelihood and non-negative matrix factorization, when testing for differential item functioning in patient-reported outcome measures using Monte Carlo simulation and real-world data. The results from this study showed that including an auxiliary variable in the process of addressing item non-response can reduce the Type I error rates. This study is a unique addition to the literature on methods to address incomplete responses in patient-reported outcome measures because it demonstrates the practicality of using observed item responses to define an auxiliary variable, which provides a basis for easy, accessible, and cost-effective approach of identifying auxiliary variable in patient-reported outcome measures.

Publication Details

Ayilara, O.F., Sajobi, T.T., Barclay, R., Jafari Jozani, M., Sawatzky, R & Lix, L.M. (2022). Enhanced weighted non-negative matrix factorization approach for addressing item non-response in patient-reported outcome measures.

6.2 Abstract

Patient-reported outcome measures (PROMs) are valuable tools to collect information about patient's perceptions of their own health in clinical trials, and population-based surveys. PROMs are prone to item non-response, which may occur for several reasons such as patients having difficulty understanding or interpreting the questions. Item non-response can invalidate group comparisons of PROM scores within the population . In this study, we propose an unsupervised machine-learning method that leverages on the aggregated observed item responses as an auxiliary variable to define weights for ordinal item-level imputation. Monte Carlo simulation and numeric example were used to compare the Type I error rates and statistical power of the proposed method, an auxiliary variable enhanced weighted non-negative matrix factorization (wNNMF-Aux) with listwise deletion, full information maximum likelihood (FIML), and non-negative matrix factorization (NNMF), when testing for differential item functioning. The simulation conditions included number of items, sample size, and percentage of missing observations. The Type I error rates and statistical power of the wNNMF-Aux method were comparable to the FIML method when the number of items was large. The wNNMF-Aux method outperformed the NNMF method for most of the simulation conditions under consideration. However, the wNNMF-Aux method resulted in inflated Type I error rates when the percentage of missing responses was more than 25% and the number of items was small. The wNNMF-Aux method has the advantage of incorporating aggregated observed item response as auxiliary variable in the imputation process to reduce Type I error rates.

6.3 Introduction

Patient-reported outcome measures (PROMs), which are comprised of questions or items that reflect underlying latent (i.e., unobserved) constructs are used in healthcare and epidemiological studies to capture information on quality of life and other aspects of health and well-being [1–3]. PROMs are used to describe disease trajectories [4] and evaluate the impact of interventions, such as behavioral therapy or surgery [5]. Item responses from PROMs instrument are multivariate and ordinal in nature (e.g., Likert scales).

PROMs are prone to item non-response (i.e., missing data), which may occur because patients have difficulty understanding or interpreting the questions, uncomfortable answering the questions due to cultural or personality differences or miss scheduled clinic visits. Item non-response can invalidate group comparisons of PROM scores within the population, and the evaluation of any clinical intervention. Item non-response may increase the risk of false-negative findings, and over- or under-estimation of differences between groups or over time [6, 7].

The choice of methods to address item non-response is dependent, in part, on the missingness mechanism, which can be missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR) [8–10]. Data are MCAR if the probability of missingness is unrelated to the observed and unobserved outcomes, that is, MCAR assumes patients with missing responses are a random sample of the study cohort or population. Data are MAR, if conditional on the observed data, the probability of missingness is independent of the unobserved outcomes. Finally, data are MNAR if the probability of missing observations depends in whole or in part, on unobserved measurements.

Ordinal item non-response methods including listwise deletion (LD) [11], multiple imputation (MI), full information maximum likelihood (FIML) and non-negative matrix factorization (NNMF) [8, 12–22] have been used to address non-response in PROMs data.

The LD method is easy to implement and widely available in most statistical software packages. Studies have shown that the LD method performs comparably to the MI, FIML and NNMF methods when testing for group differences in PROMs [11, 23, 24]. However, this method may result in biased parameter estimates and inflated standard errors, especially when the percentage of non-response is high [13].

The MI and FIML methods have been shown to perform equivalently when the imputation and analysis models are similar [19–21]. The MI method requires that the functional form of the imputation model and the distributional form of the variables included in the model are specified correctly [16, 25]. Also, potential non-linear relationships that may exist amongst variables needs to be incorporated in the imputation model [26]. The FIML method relies on a family of parametric probability model assumptions, and correct model specification [6, 27]; when these assumptions are not tenable, the FIML method may result in less than optimal performance [28].

The NNMF method makes no distributional assumption when imputing for missing response and it is computationally efficient [29]. However, the NNMF method may result in less than optimal performance when the sample size is small and the correlation among items that comprise the PROMs is weak [23].

In the context of missing responses, the optimization process of the NNMF method is equivalent to using the weighted NNMF (wNNMF) method with binary weights [30], which assigns weight

of 0 to missing response and 1 to observed response. The weighting gives the flexibility to emphasize certain item responses in the imputation process [31].

Leveraging on information from auxiliary (i.e., supplementary) variables, which are potential correlates of missingness and/or outcome of interest in missing data methods has been shown to improve model parameter estimation [20, 32–35]. Eekhout et al. showed that including aggregated observed item responses as an auxiliary variable in the FIML method increased the precision of the estimated parameters [36].

In this study, we propose an enhanced wNNMF (wNNMF-Aux) method, which uses aggregated observed item responses as an auxiliary variable to define weights before imputation; the performance of the wNNMF-Aux method was compared with the LD, FIML and NNMF methods. We do this in the context of differential item functioning (DIF), which occurs when the probability of item response varies across groups defined by such characteristics as age or sex, after adjusting for one or more latent traits represented in the items.

6.4 Methods

6.4.1 Existing Methods

In this section, we describe FIML, NNMF, and wNNMF methods.

FIML

The FIML method is a direct extension of the maximum likelihood method to missing data problem. Suppose we have a simple random sample of n individuals from the population of interest. Let $\boldsymbol{y} = (y_{i1}, y_{i2}, \dots, y_{ip})$ denote p-dimensional vector of item responses with probability density function $f(\boldsymbol{y}; \theta)$. The goal is to estimate the parameter θ. The maximum likelihood method, which chooses parameter values that assign the maximum possible

probability or probability density to the observed data under a well-defined family of parametric probability models, would be appropriate to estimate θ if $\boldsymbol{y}$ were fully observed.

Suppose only a part of $\boldsymbol{y}$ was observed for i^{th} individual, we let $(\boldsymbol{y}_i^O, \boldsymbol{y}_i^M)$ denote the subset of p items that are observed and missing for the i^{th} individual $(i = 1, \dots, n)$, respectively. Let δ_{ij} be the response indicator function of y_{ij}, with $\delta_{ij} = 1$ if y_{ij} is observed, and $\delta_{ij} = 0$ if y_{ij} is missing for $j = 1, \dots, p$ and $\boldsymbol{\delta} = (\delta_1, \dots, \delta_p)$. To estimate θ, we assume a response probability model $P(\boldsymbol{\delta}|\boldsymbol{Y}) = P(\boldsymbol{\delta}|\boldsymbol{y}; \phi)$, where ϕ is the parameter of the response probability model. For each i^{th} individual, we observe $(\boldsymbol{y}_i^O, \delta_i)$ instead of $\boldsymbol{y}_i$. The marginal density of $(\boldsymbol{y}_i^O, \delta_i)$ can be written as:

$$f(\boldsymbol{y}_i^O, \delta_i; \theta, \phi) = \int f(\boldsymbol{y}_i; \theta)\ P(\boldsymbol{\delta}_i|\boldsymbol{y}_i; \phi) d\mu(\boldsymbol{y}_i^M). \quad (1)$$

Using Eq.(1), the observed likelihood is given by:

$$L_{obs}(\theta, \phi) = \prod_{i=1}^{n} \left[\int f(\boldsymbol{y}_i; \theta)\ P(\boldsymbol{\delta}_i|\boldsymbol{y}_i; \phi) d\mu(\boldsymbol{y}_i^M) \right]. \quad (2)$$

Missing responses are integrated out from Eq.(2) to estimate the model parameter. Even under very simple distributional assumptions, the observed likelihood of the graded response model (GRM) [37, 38] have a complicated form that requires special computational techniques such as expectation-maximization (EM) algorithm [39]. Estimates obtained using the FIML method are unbiased if the missing data are MAR and the model has been correctly specified [6].

NNMF

The NNMF method approximately decomposes a non-negative matrix into the product of two low-dimensional matrices [40]. It provides a strategy for imputing missing values in a dataset

that is represented in the form of a matrix. It is often assumed that there are correlations and underlying structures in the dataset such that only a subset of observed values suffices to approximately obtain some or all missing values. The main difference between the NNMF method and other matrix factorization methods, such as singular value decomposition and maximum-margin matrix factorization, is that the two low-dimensional matrices are constrained to be non-negative. These constraints make the NNMF method applicable to address missing ordinal responses.

Suppose we have a fully observed response matrix $\boldsymbol{Y}_{p\times n}$ with low rank k $[k \ll \min(n,p)]$, where p is the number of items or questions, and n is the number of respondents. The goal of the NNMF method is to obtain an item factor matrix $\boldsymbol{U}_{p\times k}$, and a respondent factor matrix $\boldsymbol{V}_{k\times n}$, such that $\boldsymbol{Y} \approx \boldsymbol{UV}$, constraining $\boldsymbol{U}$ and $\boldsymbol{V}$ to be non-negative [40, 41]. The item factor matrix shows the relationship of each item to the latent variable, and the respondent factor matrix shows the relationship of each respondent to the latent variable.

The NNMF method involves minimizing:

$$P(\boldsymbol{U},\boldsymbol{V}) = \frac{1}{2} \|\boldsymbol{Y} - \boldsymbol{UV}\|^2, \tag{3}$$

variable.

The NNMF method involves minimizing:

$$P(\boldsymbol{U},\boldsymbol{V}) = \frac{1}{2} \|\boldsymbol{Y} - \boldsymbol{UV}\|^2, \tag{3}$$

subject to $\boldsymbol{U} \geq 0, \boldsymbol{V} \geq 0$, where $\|\,.\,\|^2$ is the squared Frobenius norm of the matrix. This optimization can be solved in different ways. We used multiplicative update rules, an iterative method to update factor matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ [42], because of their ease of implementation and

speed of convergence. The multiplicative update rules for entries u_{ik} and v_{kj} of the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ respectively are:

$$u_{ik} \Leftarrow \frac{(\boldsymbol{Y}\boldsymbol{V}^{\mathrm{T}})_{ik}u_{ik}}{(\boldsymbol{U}\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}})_{ik}}, \tag{4}$$

$$v_{kj} \Leftarrow \frac{(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{Y})_{kj}v_{kj}}{(\boldsymbol{U}^{T}\boldsymbol{U}\,\boldsymbol{V})_{kj}}. \tag{5}$$

Now, consider the situation where some entries in $\boldsymbol{Y}$ are missing. The initial formulation of the multiplicative update rules will be insufficient unless additional constraints are imposed on the rows and columns of $\boldsymbol{U}$ and $\boldsymbol{V}$. For an incomplete ordinal item response matrix $\boldsymbol{Y}$ with low rank k, let $\omega \subset \{1, \dots, p\} \times \{1, \dots, n\}$ denote the indices of observed entries, $I_j = \{i: y_{ij}$ is observed$\}$ and $\widetilde{I_j} = \{i: y_{ij}$ is missing$\}$. We observe that information in $\boldsymbol{Y}$ is redundant to achieve $\boldsymbol{Y} \approx \boldsymbol{U}\boldsymbol{V}$ since $\boldsymbol{Y}$ is assumed to have a low rank k. Therefore, factorization of the incomplete ordinal item response matrix can be achieved using the observed responses. The optimization problem in Eq.(3) can then be reformulated as,

$$\underset{U,V}{\text{minimize}}\ \frac{1}{2}\sum_{(i,j)\in\omega}(\boldsymbol{Y}_{ij} - (\boldsymbol{U}\boldsymbol{V})_{ij})^2, \tag{6}$$

subject to $\boldsymbol{U} \geq 0, \boldsymbol{V} \geq 0$. When applying the multiplicative update rules on the j^{th} column of $\boldsymbol{V}$ in Eq.(6), all $\widetilde{I_j}$ rows of $\boldsymbol{U}$ are removed. The resulting multiplicative update equation with squared Frobenius norm becomes,

$$v_{kj} \Leftarrow \frac{\left(\boldsymbol{U}_{I_{j.}}^{\mathrm{T}}\boldsymbol{Y}_{I_{j.}}\right)_{kj}v_{kj}}{\left(\boldsymbol{U}_{I_{j.}}^{\mathrm{T}}\boldsymbol{U}_{I_{j.}}\boldsymbol{V}\right)_{ik}}, \tag{7}$$

where $\boldsymbol{U}_{I_{j\cdot}}$, and $\boldsymbol{Y}_{I_{j\cdot}}$ are submatrices of $\boldsymbol{U}$ and $\boldsymbol{Y}$ with row indices in $I_{j\cdot}$ The multiplication of the resulting $\boldsymbol{U}$ and $\boldsymbol{V}$ factor matrices produce an approximation to the original data points in item response matrix $\boldsymbol{Y}$. The maximum and minimum imputed values are constrained to the highest and lowest possible item responses to avoid generation of out-of-range values. This method was implemented in RStudio version 3.5.2 using the *NNLM* package [29, 43].

wNNMF

The wNNMF method is a weighted case of the NNMF and it minimizes:

$$P_W(\boldsymbol{U},\boldsymbol{V}) = \frac{1}{2}\,\|\boldsymbol{Y} - \boldsymbol{UV}\,\|_W^2 := \frac{1}{2}\sum_{ij}[W \circ (\boldsymbol{Y} - \boldsymbol{UV}) \circ (\boldsymbol{Y} - \boldsymbol{UV})]_{ij}\,, \tag{8}$$

subject to $\boldsymbol{U} \geq 0, \boldsymbol{V} \geq 0$, where $W = \{W_{ij}\} > 0$ is a non-negative weight matrix, and $A \circ B$ is the Hadamard product (or element by element product) of the matrices *A* and *B*. Similar to the NNMF method, multiplicative update rules was used to update factor matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ [42]. We refer to Blondel et al. [31] for details on the optimization of Eq.(8). Several studies have imputed missing values using the wNNMF method by the defining a binary weights for W_{ij} as follow [30, 44]:

$$W_{ij} = \begin{cases} 1, & \text{if } \boldsymbol{Y}_{ij} \text{ is observed} \\ 0, & \text{if } \boldsymbol{Y}_{ij} \text{ is unobserved.} \end{cases}$$

When all the elements in $W_{ij} = 1$, Eq.(8) reduces to Eq.(3). This method was implemented in RStudio version 3.5.2 using the *wNMF* function in Python [31].

6.4.2 Proposed Method

wNNMF-Aux

In this section we introduce the wNNMF-Aux method to address ordinal item non-response. This method is an extension of the wNNMF method. The framework of our proposed method is presented in Algorithm 1.

Algorithm 1: The wNNMF-Aux method algorithm

1. Define an auxiliary variable *Z*, where *Z* is the sum of the observed item responses for the i^{th} individual in the incomplete ordinal item response matrix $\boldsymbol{Y}$.
2. Partition $\boldsymbol{Y}$ into *q* distinct, non-overlapping groups based on the estimated quantiles of *Z*.
3. Initialize the missing responses for i^{th} individual with the median of the observed responses on the corresponding items.
4. Define W_{ij} within each group as the inverse of the ratio of the number of individuals with at least one missing response to the group size.
5. Minimize Eq.(8) as before with the newly defined weights and initialized missing responses

$$P_W(\boldsymbol{U},\boldsymbol{V}) = \frac{1}{2}\sum_{ij}[W \circ (\boldsymbol{Y} - \boldsymbol{UV}) \circ (\boldsymbol{Y} - \boldsymbol{UV})]_{ij},$$

subject to $\boldsymbol{U} \geq 0, \boldsymbol{V} \geq 0$.

6. Replace the missing response for i^{th} individual with predicted response

6.4.3 Simulation Study

Data generation and simulation conditions

We simulated unidimensional polytomous data for *p* items (*p* = 15 and 30) using GRM:

$$P\left(\boldsymbol{Y}_{ij} \geq c \mid \boldsymbol{\theta}, \alpha_j, \beta_{jc}\right) = \frac{exp\left[\alpha_j\left(\theta_i - \beta_{jc}\right)\right]}{1 + exp\left[\alpha_j\left(\theta_i - \beta_{jc}\right)\right]}, \quad (9)$$

where $P\left(\boldsymbol{Y}_{ij} \geq c \mid \boldsymbol{\theta}, \alpha_j, \beta_{jc}\right)$ is the probability of selecting the response option c $(c = 1, \cdots, C)$ or higher on item j $(j = 1, \cdots, p)$ by respondent i $(i = 1, \cdots, n)$, with latent trait θ, α_j is the slope of item j, and β_{jc} is the threshold of item j for the c^{th} category.

The slopes ($\boldsymbol{\alpha}$) of the items were sampled from a uniform distribution $U(1.1, 2.8)$ [45]. The thresholds ($\boldsymbol{\beta}$) also followed a uniform distribution $U(-2, 2)$ [46, 47]. The number of item response categories (C) was set to $C = 5$, and the latent trait ($\boldsymbol{\theta}$) followed a normal distribution, $N(0, 1)$.

Missing data were simulated for the items that exhibited DIF [24, 48, 49]. The missingness on these items was achieved using the multivariate amputation method [50] that generates missing data by assigning a weight to each observation. For example, to achieve missingness under the MAR mechanism, the item weights depend only on the observed data; the probability of missingness is calculated assuming a logistic distribution based on the weights. Observations with a larger weight have a greater likelihood of missingness. This method has been shown to be more appropriate for generating missing data when compared to a stepwise univariate amputation method (i.e., missingness on items are generated one at a time) [51, 52]. Pre-specified amounts of responses were removed from the DIF items for each missing data mechanism using the multivariate *ampute* function in the *mice* package [53] in *R version 3.5.2*.

The simulation conditions including sample size, percentage of missing data, magnitude of uniform and non-uniform DIF effects are presented in Table 1. Random samples of sizes (n) were generated for each of the reference (R) and focal (F) groups and denoted as R500/F500, R1000/F1000, and R1500/F500 giving total sample sizes that ranged from 1000 to 2000. The fully observed condition (i.e., 0% missing responses) was used as a reference to assess the performance of item non-response methods. The magnitude of the difference between the groups' thresholds and slopes were set to 0 (i.e., no DIF present), 0.2 (small), 0.4 (moderate) and 0.6 (large) [54]. A total of 500 replications were conducted for each of the 162 combinations of simulation conditions.

Table 6-1. Manipulated simulation conditions

Conditions	**Values**
Number of items items (*p*)	15, 30
Random samples of size (*n*)	R500/F500, R1000/F1000, R1500/F500
Pre-specified % of missing responses	0%, 10%, 25%, 50%
Missingness mechanisms	MCAR, MAR, MNAR
% of items that exhibited DIF effects	20%
Magnitude of DIF effects	0, 0.2, 0.4, 0.6

R: Reference group; F: Focal group

Simulation scenarios

We considered a number of simulation scenarios to compute the Type I error rates and power of the test. In the first scenario, which assumes neither uniform nor non-uniform DIF were present in the simulated data, no DIF effect was added to the threshold and slope parameters of the DIF target items. In the second scenario, we considered two conditions where DIF was present. Under the first condition where only uniform DIF was present, we added small to large DIF effects to the threshold parameters of the DIF target items in the focal group. Under the second condition where non-uniform DIF was present, we added small to large DIF effects to the slope parameters of the DIF target items in the focal group.

We used the item response theory likelihood ratio test (IRT-LRT) to test for uniform and non-uniform DIF because of its popularity and availability in statistical software packages such as RStudio. The IRT-LRT compares the difference in the likelihood between two nested GRMs using a χ^2 statistic, with degrees of freedom equal to the difference in degrees of freedom between the two GRMs [48].

To test for DIF when it was not present, we fit a DIF model that assumed all items, except the target items were invariant between the focal and reference groups. Next, we fit a set of no-DIF

models that constrained the threshold or slope parameters to be equal for all items in the two groups. A statistically significant LRT indicated the presence of DIF.

To test for uniform or non-uniform DIF when it was present, we fit a DIF model as specified above. Next, we fit no-DIF models that constrained the threshold or slope parameters to be equal for all items in the two groups. A statistically significant LRT when the DIF model was compared to: (1) a no-DIF model that constrained the threshold parameters for all items in the two groups indicated the presence of uniform DIF, and (2) a no-DIF model, where the slope parameters were constrained for all items in the two groups indicated the presence of non-uniform DIF. The GRM-LRT was implemented in RStudio version 3.5.2. using the *mirt* package [55].

Performance measures

The performance of the item non-response methods was evaluated using the power of the test and Type I error rate. We report the average percentage of statistically significant LRT statistics across replications and items that were tested for DIF.

Bradley's liberal criterion (i.e., 0.025 – 0.075) for $\alpha = 0.05$ was used to evaluate the robustness of Type I error rates [56]. The power of the GRM-LRT with and without adoption of a missing data method was compared; differences in the power of the LRT ≤10% were considered insubstantial while differences >10% were considered substantial [57, 58].

6.4.4 Numeric Example

Data source and measures

The computerized adaptive test (CAT-5D-QOL), which consists of five domains that are related to joint problems (i.e., pain or discomfort, daily activities, walking, handling objects, and

feelings) was used as a real-world example to demonstrate the implementation of the item non-response methods. Details of the methodology to develop and collect the study data for the CAT-5D-QOL have been described elsewhere [59, 60].

We limit our attention to 35 out of the 36 items measuring the severity and frequency of pain or discomfort (PAIN) and the impact of pain on activities of daily living and leisure for this demonstration. The response options varied for the items; 16 items had responses that ranged from 1 to 5 for “not at all” to “extremely”, 12 items had responses that ranged from 1 to 5 for “never” to “always”, and 7 items had a range of item-specific response options. The reliability and validity of the CAT-5D-QOL has been reported in other studies [59–61]. Consistent with previous studies, we test for DIF by sex [62].

Statistical analyses

Descriptive statistics were used to describe the study cohort. A three-step procedure was used to test for uniform and non-uniform DIF using the GRM with the LRT to: (1) assess the dimensionality of the CAT-5D-QOL PAIN domain, (2) select anchor items, and (3) assess DIF [63] (see Supplementary Material 1).

Exploratory and confirmatory analyses were used to assess the dimensionality of CAT-5D-QOL PAIN domain [64, 65]. The exploratory approach compared the first and second eigenvalues obtained from the polychoric correlations for all the items. The first eigenvalue should be significantly greater than the second eigenvalue (i.e., ratio > 4:1) to support unidimensionality [64]. The confirmatory approach compared the comparative fit index (CFI) root mean square error of approximation (RMSEA), and standardized root mean squared residual (SRMR) of unidimensional and two-dimensional graded response models using a unit variance identification

constraint (i.e., constraining the mean and variance of each factor to 0 and 1 respectively). A RMSEA ≤ 0.05, a SRMR ≤ 0.08, and a CFI > 0.95 indicate close or good model fit [66, 67]. Items with complete responses was used to assess the dimensionality.

Model parameters were estimated using the maximum likelihood method. The Hochberg step-up approach was used to control the familywise Type I error rate to $\alpha = 0.05$ [68].

6.5 Results

6.5.1 Simulation Study Results

Fully observed item responses

Table 2 shows the Type I error and power rates for the LRTs of uniform and non-uniform DIF when the data were fully observed (i.e., no missing responses). The Type I error rates for all the simulation conditions considered were within the bounds of Bradley's liberal criterion [56]. The number of items, sample sizes and magnitude of DIF effects were positively associated with statistical power. For uniform DIF, power ranged from 21.2% to 100.0% and for non-uniform DIF, power ranged from 17.5% to 96.4%.

Table 6-2. Type I error rates (%) and power rates (%) to detect uniform and non-uniform DIF in fully observed data for the likelihood ratio test

# of items	Sample size	Type I error	Power to detect uniform DIF			Power to detect non-uniform DIF		
			Magnitude of DIF effect					
			0.2	0.4	0.6	0.2	0.4	0.6
15	R1000/F1000	5.3	34.6	93.4	100.0	25.8	67.9	91.2
	R1500/F500	5.0	26.3	84.8	99.7	22.1	58.1	86.0
	R500/F500	6.3	21.2	68.1	96.2	17.8	44.4	69.7
30	R1000/F1000	4.6	39.3	96.0	100.0	30.9	75.2	94.5
	R1500/F500	4.8	28.5	87.1	99.9	23.4	63.5	89.0
	R500/F500	5.5	21.2	69.4	96.8	17.5	47.6	75.6

R: Reference group; F: Focal group

Missing item responses

The Type I error percentages for detecting DIF using the LRT when items contained missing responses are reported in Table 3. The results are stratified by number of items, sample size and percentage of missing responses. The Type I error rates for the wNNMF-Aux method were within the bounds of the liberal criterion for more than 95.0% of the conditions, but this method did exhibit inflated error rates as the number of items decreased. The error rates of the NNMF method exceeded the upper bound of Bradley's liberal criterion, indicating inflated Type I error rates, especially with 50% missing responses. The percentages for the LD and FIML methods were within the bounds of the liberal criterion for all the investigated simulation conditions.

Table 6-3. Type I error rates (%) for item non-response methods when testing for DIF using the likelihood ratio test

# of items	Mech.	Sample size	Percentage of missing responses											
			10%				25%				50%			
			LD	FIML	NNMF	wNNMF-Aux	LD	FIML	NNMF	wNNMF-Aux	LD	FIML	NNMF	wNNMF-Aux
15	MCAR	R1000/F1000	4.8	4.9	5.1	4.8	5.2	5.3	5.8	5.7	5.0	4.9	**7.8**	7.3
		R1500/F500	4.8	4.7	4.3	4.3	5.3	5.2	5.5	5.4	5.0	5.0	7.2	7.2
		R500/F500	5.3	5.3	5.8	6.1	5.3	5.2	5.3	5.4	5.7	5.4	**8.2**	**7.8**
	MAR	R1000/F1000	5.4	5.1	5.3	5.4	5.6	5.6	6.3	5.7	4.7	4.8	**8.3**	**8.4**
		R1500/F500	5.6	5.6	5.9	5.8	5.1	5.1	6.7	5.7	5.8	5.8	**8.4**	**8.1**
		R500/F500	5.6	5.7	6.2	6.3	5.9	6.1	7.0	6.3	5.1	5.2	**8.4**	**8.0**
	MNAR	R1000/F1000	5.3	5.3	5.0	5.4	6.0	5.8	6.4	5.9	4.9	5.2	**7.9**	**7.9**
		R1500/F500	5.6	5.5	5.8	5.7	5.7	5.8	6.6	6.5	5.7	5.7	**8.0**	**7.8**
		R500/F500	5.8	5.7	6.3	6.1	5.8	5.8	7.4	6.9	5.6	5.7	**8.2**	**7.7**
30	MCAR	R1000/F1000	4.3	4.3	4.8	4.8	4.7	4.7	5.9	5.7	5.1	5.0	**7.7**	7.1
		R1500/F500	5.0	4.9	5.4	5.0	4.8	4.8	5.6	5.4	4.5	4.5	**7.8**	7.2
		R500/F500	5.4	5.4	5.3	5.3	4.8	4.8	6.1	6.1	5.7	5.7	6.9	7.0
	MAR	R1000/F1000	4.7	4.6	4.9	4.8	4.1	4.2	5.2	5.0	4.6	4.6	**7.6**	7.0
		R1500/F500	4.5	4.5	4.1	4.6	4.5	4.5	5.2	5.1	5.0	5.0	**7.7**	7.2
		R500/F500	5.4	5.3	5.7	5.7	5.3	5.2	6.0	5.8	4.8	4.8	**7.9**	7.3
	MNAR	R1000/F1000	4.7	4.6	5.3	5.0	4.4	4.4	5.0	5.0	4.7	4.6	**7.8**	7.3
		R1500/F500	4.5	4.5	4.3	4.5	4.1	4.1	4.8	4.7	5.1	5.2	7.1	7.3
		R500/F500	5.3	5.4	5.6	5.5	4.9	4.9	6.4	6.4	5.1	5.1	**7.8**	7.4

MCAR: Missing completely at random; MAR: Missing at random; MNAR: Missing not at random; LD: List-wise deletion; FIML: Full information maximum likelihood; NNMF: Non-negative matrix factorization; wNNMF-Aux: Enhanced weighted NNMF; R: Reference group; F: Focal group; Boldface font indicates error rates that fall outside the bounds of Bradley's [56] liberal criterion (i.e., 2.5 – 7.5%)

The difference in power of the LRT to detect uniform DIF between fully observed and missing data methods for 15 and 30 items, and sample size of R1000/F1000 stratified by the magnitude of DIF and the missing data mechanism for all item non-response methods are presented in Figures 1 and 2, respectively.

The power to detect uniform DIF when 50% of the responses were missing and sample size was set to R1000/F1000, ranged between 18.1% and 97.3% for LD; for FIML the corresponding values were 17.1% and 97.5%; for NNMF the corresponding values were 20.6% and 96.3%; and for wNNMF-Aux the corresponding values were 19.6% and 96.4%. The power to detect non-uniform DIF when 50% of the responses were missing for sample size of R1000/F1000 ranged between 13.2% and 77.7% for the LD method; for the FIML method the corresponding values were 13.2% and 77.4%; for NNMF the corresponding values were 19.1% and 71.7%; and for wNNMF-Aux the corresponding values were 20.2% and 73.0%.

When the number of items was set to 15 and sample size was set to R1000/F1000, the power to detect DIF effects (uniform and non-uniform) ranged between 13.2% and 100.0% for LD and FIML; for NNMF the corresponding values were 19.1% and 100.0%; for wNNMF-Aux the corresponding values were 20.2% and 100.0%. The power to detect DIF effects when number of items was set to 30 and sample size set to R1000/F1000 ranged between 15.0% and 99.9% for the LD method; for the FIML method the corresponding values were 14.8% and 99.9%; for NNMF the corresponding values were 20.5% and 99.9%; and for wNNMF-Aux the corresponding values were 20.9% and 99.9%.

The differences in power to detect small (i.e., 0.2) and moderate (i.e., 0.4) uniform DIF effects between fully observed and item non-response methods when the percentage of non-response

was 50% were substantial for all the methods. The pattern of power results was similar for sample sizes of R1500/F500 and R500/F500 (see Supplementary Material 2).

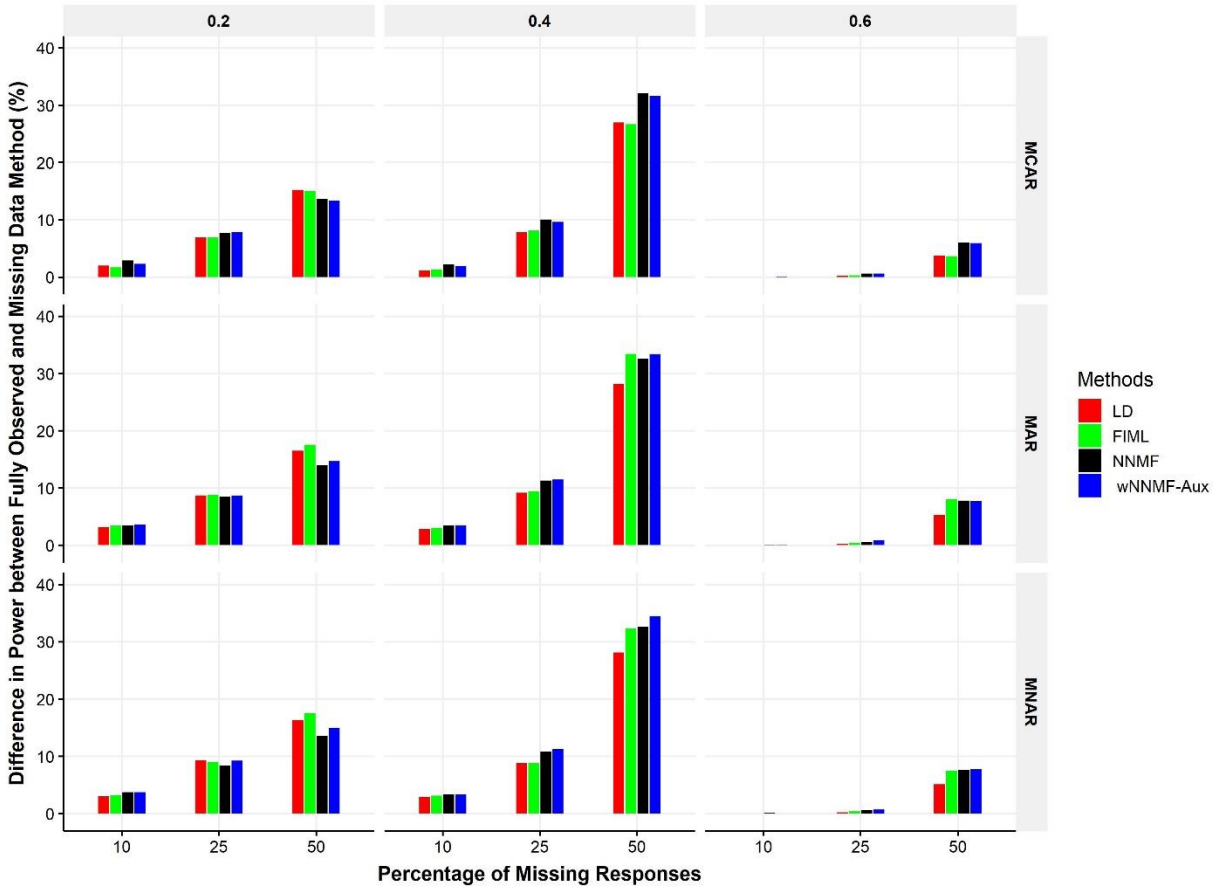


Figure 6-1. Difference in Power of the LRT to detect **uniform DIF** between fully observed and missing data method (%) for **15 items**, and sample size of R1000/F1000 (R: reference group; F: focal group) stratified by the missing data mechanism and magnitude of DIF. MCAR: missing completely at random; MAR: missing at random; and MNAR: missing not at random.

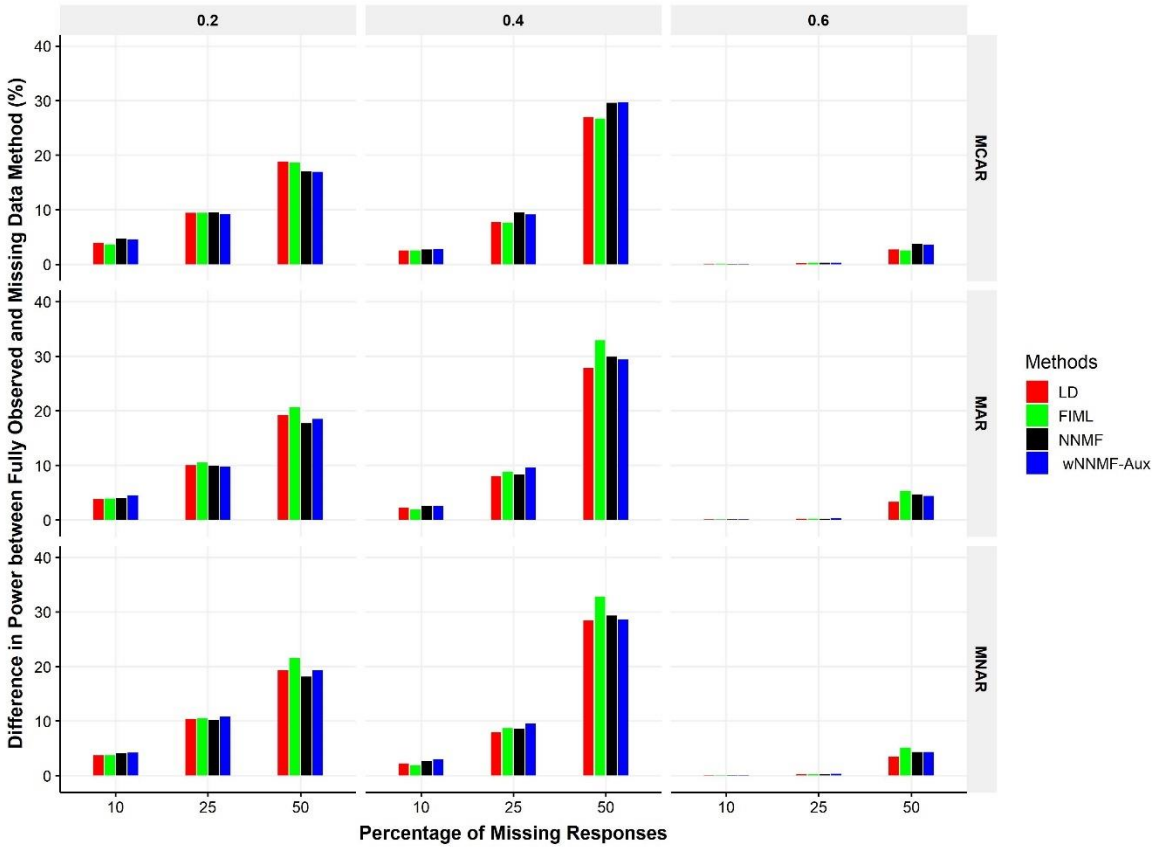


Figure 6-2. Difference in Power of the LRT to detect **uniform DIF** between fully observed and missing data method (%) for **30 items**, and sample size of R1000/F1000 (R: reference group; F: focal group) stratified by the missing data mechanism and magnitude of DIF. MCAR: missing completely at random; MAR: missing at random; and MNAR: missing not at random.

6.5.2 Numeric Example Results

The study cohort of the real-world data was comprised of 1660 patients. More than half (59.6%) were women and the average age was 56.6 years old.

The distribution of responses of the items for the PAIN domain are shown in Supplementary Table 1. The average missing response rate per item (i.e., the sum of the percentage of missing responses for each item divided by the total number of items) was 1.2%. Approximately 10.0% of the total sample size had at least one missing response.

The first and second eigenvalues from the exploratory analysis had values of 26.10 and 1.98 respectively. The ratio of the eigenvalues was 13.17, which suggests unidimensional latent variable structure. The confirmatory analysis showed that a unidimensional model had a reasonable fit to the data (RMSEA = 0.08, 90% CI = 0.084 – 0.088; CFI = 0.95; SRMR = 0.06). Item Q173 (pain restrict from routine daily activities) had the highest slopes amongst all items that were non-significant using the all-other anchor approach (AOAA) (69,70). Therefore, this item was selected as anchor item.

Table 4 shows the difference in the LRT statistics for unconstrained and constrained DIF models of the items on the PAIN domain for each of the item non-response methods. None of the PAIN items showed any evidence of uniform and non-uniform DIF.

Table 6-4. Difference in the log-likelihood statistics of an unconstrained and constrained DIF model by sex (degrees of freedom) for the CAT-5D-QOL PAIN items in the real-world data for the item non-response methods

Item	**Uniform DIF**				**Non-uniform DIF**			
PAIN	**LD**	**FIML**	**NNMF**	**wNNMF-Aux**	**LD**	**FIML**	**NNMF**	**wNNMF-Aux**
Q138: Select one that describes you best	1.13 (3)	1.73 (3)	2.63 (3)	2.74 (3)	-0.06 (1)	0.18 (1)	0.08 (1)	0.06 (1)
Q139: How much bodily pain have you had?	4.14 (4)	3.98 (4)	6.67 (4)	6.49 (4)	2.82 (1)	2.01 (1)	1.39 (1)	1.46 (1)
Q140: Pain interference in a number of activities	8.05 (3)	6.00 (3)	3.83 (3)	4.00 (3)	0.96 (1)	0.41 (1)	0.48 (1)	0.72 (1)
Q141: Pain interference with normal work	4.58 (4)	4.22 (4)	10.24 (4)	9.03 (4)	1.95 (1)	1.79 (1)	6.24 (1)	5.71 (1)
Q142: Pain affect ability to fall asleep	3.25 (4)	2.56 (4)	1.93 (4)	2.13 (4)	0.33 (1)	0.29 (1)	0.42 (1)	0.49 (1)
Q143: Pain interference with physical leisure	2.75 (4)	2.79 (4)	8.61 (4)	7.71 (4)	0.24 (1)	0.25 (1)	1.35 (1)	1.05 (1)
Q144: Intense pain	5.21 (4)	3.96 (4)	7.41 (4)	7.07 (4)	1.44 (1)	-0.01 (1)	0.86 (1)	0.95 (1)
Q146: Pain interference with non-physical leisure	5.72 (4)	4.08 (4)	4.68 (4)	4.70 (4)	0.09 (1)	-0.08 (1)	0.05 (1)	0.08 (1)
Q147: Pain interference with social activities	1.99 (4)	2.94 (4)	2.79 (4)	2.69 (4)	0.23 (1)	0.23 (1)	0.95 (1)	1.06 (1)
Q148: Pain prevent from taking bath or shower	10.07 (4)	5.93 (4)	1.94 (4)	1.72 (4)	1.04 (1)	0.65 (1)	0.24 (1)	0.28 (1)
Q149: Pain interference with self-care activities	5.96 (4)	5.15 (4)	2.26 (4)	2.54 (4)	0.05 (1)	0.14 (1)	0.45 (1)	0.56 (1)
Q150: Pain affect ability to use the toilet	3.07 (3)	4.89 (3)	1.97 (3)	1.94 (3)	0.46 (1)	0.23 (1)	3.06 (1)	2.85 (1)
Q151: Pain interference with light household chores	7.48 (4)	6.69 (4)	5.69 (4)	5.27 (4)	2.82 (1)	2.58 (1)	0.27 (1)	0.20 (1)
Q152: Pain prevent from dressing and undressing	6.52 (4)	4.38 (4)	1.28 (4)	1.34 (4)	2.35 (1)	3.35 (1)	0.27 (1)	0.27 (1)
Q153: Pain prevent from using the toilet	2.39 (2)	2.08 (2)	1.14 (2)	0.91 (2)	2.98 (1)	1.73 (1)	0.00 (1)	0.00 (1)
Q154: Pain interference with physical leisure	4.78 (4)	5.72 (4)	6.54 (4)	6.65 (4)	-0.04 (1)	-0.04 (1)	0.85 (1)	0.94 (1)
Q155: Pain prevent from grooming	8.68 (3)	3.46 (3)	0.82 (3)	0.71 (3)	-0.06 (1)	-0.03 (1)	0.00 (1)	0.00 (1)
Q157: Pain interference with sleep	2.42 (4)	1.65 (4)	3.86 (4)	3.88 (4)	0.06 (1)	0.05 (1)	2.64 (1)	2.83 (1)
Q158: Pain interference with social activities outside the home	1.72 (4)	3.00 (4)	3.81 (4)	4.15 (4)	0.03 (1)	-0.06 (1)	0.42 (1)	0.38 (1)
Q159: Pain interference with eating meals	2.55 (3)	3.24 (3)	0.44 (3)	0.41 (3)	-0.04 (1)	-0.02 (1)	0.01 (1)	0.01 (1)
Q160: Pain affect ability to accomplish more	3.26 (4)	4.77 (4)	1.12 (4)	1.17 (4)	-0.06 (1)	0.09 (1)	0.67 (1)	0.62 (1)
Q164: Feel great physically	13.40 (4)	11.98 (4)	3.83 (4)	3.40 (4)	9.50 (1)	6.58 (1)	0.00 (1)	0.03 (1)
Q166: Pain prevent from eating meal	3.63 (3)	4.07 (3)	2.82 (3)	2.83 (3)	0.50 (1)	0.28 (1)	0.08 (1)	0.10 (1)
Q167: Pain restrict from routine daily activities	2.33 (4)	3.67 (4)	3.78 (4)	3.75 (4)	-0.06 (1)	-0.03 (1)	2.21 (1)	2.59 (1)

Q168: Avoid strenuous activities for fear of pain	5.48 (4)	3.75 (4)	6.06 (4)	6.09 (4)	0.35 (1)	0.97 (1)	1.87 (1)	2.19 (1)
Q169: Pain limitation in doing things	5.20 (4)	3.96 (4)	4.57 (4)	4.39 (4)	1.77 (1)	0.80 (1)	1.24 (1)	1.47 (1)
Q170: Pain frustration	5.21 (4)	5.23 (4)	4.90 (4)	4.80 (4)	1.75 (1)	3.13 (1)	3.11 (1)	3.29 (1)
Q171: Pain prevent from routine daily activities	7.71 (4)	7.59 (4)	2.52 (4)	2.19 (4)	0.64 (1)	0.59 (1)	2.18 (1)	2.58 (1)
Q172: Wake up at night due to pain	12.08 (4)	11.33 (4)	4.05 (4)	3.61 (4)	0.50 (1)	0.58 (1)	1.26 (1)	1.49 (1)
Q173: Pain restrict from routine daily activities	†	†	†	†	†	†	†	†
Q174: Which one best describes level of pain?	6.64 (4)	6.12 (4)	4.15 (4)	4.14 (4)	2.59 (1)	1.74 (1)	0.88 (1)	0.87 (1)
Q175: Help in dressing or bathing due to pain	0.70 (4)	1.82 (4)	0.43 (4)	0.43 (4)	3.06 (1)	2.51 (1)	0.00 (1)	0.01 (1)
Q176: Stay in bed because of pain	0.77 (3)	0.46 (3)	7.07 (3)	7.00 (3)	1.35 (1)	1.99 (1)	2.48 (1)	2.64 (1)
Q177: Unbearable pain	1.27 (3)	1.28 (3)	1.52 (3)	2.09 (3)	0.32 (1)	-0.02 (1)	0.13 (1)	0.12 (1)
Q178: Which statement best describes your pain?	2.52 (2)	4.70 (2)	0.46 (2)	0.49 (2)	0.11 (1)	-0.05 (1)	0.25 (1)	0.06 (1)

†= anchor item; LD: List-wise deletion; FIML: full information maximum likelihood; NNMF: Non-negative matrix factorization; wNNMF-Aux: Enhanced Weighted Non-negative matrix factorization

6.6 Discussion

This study proposed an enhanced wNNMF-Aux method for addressing item non-response and compared the performance to the LD, FIML and NNMF methods when testing for DIF in PROMs via simulated and real-world data. The wNNMF-Aux method has the advantage of leveraging on information from auxiliary variable that may be related with missingness in its imputation process.

The results indicate that the wNNMF-Aux method showed comparable performance to the LD and FIML methods but outperformed the NNMF method for most of the simulation conditions under consideration, especially when the number of items was large. The Type I error rates for the wNNMF-Aux, LD, FIML and NNMF methods were within the bounds of Bradley's liberal criterion [56] with 10% missing responses. When the percentage of missing responses was more than 25% and the number of items was small, the wNNMF-Aux method resulted in inflated error rates.

The Type I error rates for the wNNMF-Aux method were within the bounds of Bradley's liberal criterion [56] for more than 95% of the investigated simulation conditions. While the Type I error rates for the LD and FIML methods were within the bounds of Bradley's liberal criterion for all the investigated simulation conditions. These findings for the LD and FIML methods are consistent with previous studies [23, 24, 49].

Power to detect uniform and non-uniform DIF increased as sample size, number of items and magnitude of DIF effects increased but decreased as the percentage of item non-response increased across all methods. This finding is also consistent with previous studies [24, 71, 72]. The differences in power to detect moderate DIF effects for the LD and FIML, wNNMF-Aux methods were substantial with 50% missing responses and insubstantial for less than 50%

missing responses, regardless of the number of items. However, the differences in power to detect moderate DIF effects for the NNMF and wNNMF-Aux methods were substantial with 50% missing responses and insubstantial for less than 50% missing responses when the number of items was large.

As expected, the performance of all the item non-response methods under consideration were similar in the analysis of the CAT-5D-QOL data. This is because the percentage of missing responses in the CAT-5D-QOL data was approximately 10%. This finding is consistent with the conclusion from our simulation results for 10% missing responses. There was no evidence of DIF for sex in any of the items on the PAIN domain, which is consistent with previous research [62].

6.6.1 Strengths and Limitations

This study demonstrated the practicality of using observed item responses to define an auxiliary variable, which provides a basis for easy, accessible, and cost-effective approach of identifying auxiliary variable in PROMs. We used a combination of simulation and real-world data to investigate the performance of our proposed method. Several data conditions were explored via the simulation study, while practical example was illustrated through the real-world data.

There are, however, some limitations to this study. First, like other simulation studies, the simulation conditions we investigated were not exhaustive. We considered one auxiliary variable, larger numbers for this condition would substantially increase computational time. Total sample sizes ranged between 1000 and 2000, smaller sample sizes may result in model convergence problem [55]. The item non-response methods we compared to the wNNMF-Aux method are best suited for MCAR and MAR cases. Therefore, a future study should compare the

performance of the wNNMF-Aux method with methods that are designed to address MNAR missingness mechanism.

6.6.2 Conclusion

In summary, we proposed an enhanced wNNMF-Aux method and compared it with different item non-response methods when testing for uniform and non-uniform DIF in PROMs data. We found that the wNNMF-Aux method outperformed the NNMF method in most of the simulation conditions considered. Hence, including an auxiliary variable in the model to address item non-response reduced the Type I error rates. Overall, the wNNMF-Aux method compared favorably to the LD and FIML methods, although it did occasionally result in inflated Type I error rates as the number of items decreased. Statistical power for all the methods were reasonably close to those from fully observed data when testing for large DIF effect.

There was no single method that performed best in terms of statistical power. The wNNMF-Aux method has the advantage of incorporating information from auxiliary variable that may be associated with missingness. Given the potential benefits of auxiliary variable information in addressing item non-response, we recommend including aggregated observed item responses as an auxiliary variable in the imputation process when addressing missing responses.

Declarations

Ethics approval and consent to participate. Not applicable

Consent for publication. Not applicable.

Availability of data and material. Not applicable.

Competing interests. The authors declare that they have no competing interests.

Funding. Funding for this study was provided by the Canadian Institutes of Health Research (grant # MOP-142404). OFA is supported by funding from the Visual and Automated Disease Analytics (VADA) Program at the University of Manitoba. LML was supported by a Tier 1 Canada Research Chair in Methods for Electronic Health Data Quality. MJJ acknowledges the Natural Sciences and Engineering Research Council of Canada for partial support. RB acknowledges the research support of the Canadian Institutes of Health Research.

Authors' contributions. All authors conceived the study and prepared the analysis plan. OFA, MJJ and LML conducted the analysis and prepared the draft manuscript. All authors reviewed and approved the final version of the manuscript.

6.7 References

1. Berzon R, Hays RD, Shumaker SA. International use, application and performance of health-related quality of life instruments. Quality of Life Research. 1993 Dec;2(6):367–8.
2. Deshpande PR, Rajan S, Sudeepthi BL, Abdul Nazir CP. Patient-reported outcomes: a new era in clinical research. Perspectives in Clinical Research. 2011 Oct;2(4):137–44.
3. Devji T, Johnston BC, Patrick DL, Bhandari M, Thabane L, Guyatt GH. Presentation approaches for enhancing interpretability of patient-reported outcomes (PROs) in meta-analysis: a protocol for a systematic survey of Cochrane reviews. BMJ Open. 2017;7(9):e017138.
4. Estabrook R, Cella D, Zhao F, Manola J, DiPaola RS, Wagner LI, et al. Longitudinal and dynamic measurement invariance of the FACIT-fatigue scale: an application of the measurement model of derivatives to ECOG-ACRIN study E2805. Quality of Life Research. 2018 Jun;27(6):1589–97.
5. Jabrayilov R, Emons WHM, de Jong K, Sijtsma K. Longitudinal measurement invariance of the Dutch outcome questionnaire-45 in a clinical sample. Quality of Life Research. 2017;26(6):1473–81.
6. Bell ML, Fairclough DL. Practical and statistical issues in missing data for longitudinal patient-reported outcomes. Statistical Methods in Medical Research. 2014;23(5):440–59.
7. Little RJA. Modeling the drop-out mechanism in repeated-measures studies. American Statistical Association. 1995;90(431):1112–21.
8. Peyre H, Leplège A, Coste J. Missing data methods for dealing with missing items in quality of life questionnaires. A comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French . Quality of Life Research. 2011;20(2):287–300.
9. Myers WR. Handling missing data in clinical trials: an overview. Drug Information Journal. 2000;34:525–33.
10. Little RJA, Rubin DB. Statistical analysis with missing data. Second Edi. New York: Wiley; 2002. 1–409 p.
11. Banks K. An introduction to missing data in the context of differential item functioning. Practical Assessment, Research and Evaluation. 2015;20(12):1–10.
12. Wu W, Jia F, Enders C. A comparison of imputation strategies for ordinal missing data on Likert scale variables. Multivariate Behavioral Research. 2015;50(5):484–503.
13. Eekhout I, De Vet HCW, Twisk JWR, Brand JPL, De Boer MR, Heymans MW. Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. Journal of Clinical Epidemiology. 2014;67(3):335–42.

14. Donneau AF, Mauer M, Molenberghs G, Albert A. A simulation study comparing multiple imputation methods for incomplete longitudinal ordinal data. Communications in Statistics: Simulation and Computation. 2015;44(5):1311–38.

15. Sulis I, Porcu M. Handling missing data in item response theory. Assessing the accuracy of a multiple imputation procedure based on latent class analysis. Journal of Classification. 2017;34(2):327–59.

16. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research. 2007;16(3):219–42.

17. Jia F, Wu W. Evaluating methods for handling missing ordinal data in structural equation modeling. Behavior Research Methods. 2019;51(5):2337–55.

18. Finch WH. Imputation methods for missing categorical questionnaire data: a comparison of approaches. Journal of Data Science. 2010;8(3):361–78.

19. Ayilara OF, Zhang L, Sajobi TT, Sawatzky R, Bohm E, Lix LM. Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. Health and Quality of Life Outcomes. 2019;17(1):106.

20. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods. 2001;6(4):330–51.

21. Schafer JL, Graham JW. Missing data: Our view of the state of the art. Psychological Methods. 2002;7(2):147–77.

22. Ayilara OF, Sajobi TT, Barclay R, Jafari Jozani M, Lix LM. Estimating longitudinal change in latent variable means: a comparison of non-negative matrix factorization and other item non-response methods. Journal of Statistical Computation and Simulation. 2022;1–20.

23. Ayilara OF, Sajobi TT, Barclay R, Bohm E, Jozani Jafari M, Lix LM. A comparison of methods to address item non-response when testing for differential item functioning in multidimensional patient-reported outcome measures. Quality of Life Research. 2022;1–12.

24. Finch H. The use of multiple imputation for missing data in uniform DIF analysis: Power and Type I error rates. Applied Measurement in Education. 2011;24(4):281–301.

25. Schafer JL, Olsen MK. Multiple imputation for multivariate missing-data problems: a data analyst's perspective. Multivariate Behavioral Research. 1998;33(4):545–71.

26. Nguyen CD, Carlin JB, Lee KJ. Model checking in multiple imputation: an overview and case study. Emerging Themes in Epidemiology. 2017;14(1):1–12.

27. Donneau AF, Mauer M, Lambert P, Molenberghs G, Albert A. Simulation-based study comparing multiple imputation methods for non-monotone missing ordinal data in longitudinal settings. Journal of Biopharmaceutical Statistics. 2015;25(3):570–601.

28. Baker FB, Kim SH. Item response theory: parameter estimation techniques. 2nd ed. Baker FB, Kim SH, editors. Boca Raton: CRC Press; 2004. 528 p.

29. Lin XE, Boutros PC. Optimization and expansion of non-negative matrix factorization. BMC Bioinformatics. 2020;21(1):1–10.

30. Kim Y, Choi S. Weighted nonnegative matrix factorization. In: IEEE International Conference on Acoustics, Speech and Signal Processing . IEEE; 2009. p. 1541–4.

31. Blondel VD, Ho ND, van Dooren P. Weighted nonnegative matrix factorization and face feature extraction. 2007.

32. Mustillo S, Kwon S. Auxiliary variables in multiple imputation when data are missing not at random. The Journal of Mathematical Sociology. 2015;39(2):73–91.

33. Ibrahim JG, Lipsitz SR, Horton N. Using auxiliary data for parameter estimation with non-ignorably missing outcomes. Applied Statistics. 2001;50(3):361–73.

34. Yoo JE. The effect of auxiliary variables and multiple imputation on parameter estimation in confirmatory factor analysis. Educational and Psychological Measurement. 2009;69(6):929–47.

35. Enders CK. A note on the use of missing auxiliary variables in full information maximum likelihood-based structural equation models. Structural Equation Modeling. 2008;15(3):434–48.

36. Eekhout I, Enders CK, Twisk JWR, de Boer MR, de Vet HCW, Heymans MW. Analyzing incomplete item scores in longitudinal data by including item score information as auxiliary variables. Structural Equation Modeling. 2015;22(4):588–602.

37. Samejima F. Graded response model. In: van der Linden WJ, Hambleton RK, editors. Handbook of Modern Item Response Theory. New York: Springer; 1997. p. 85–100.

38. Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika. 1969;34(S1):1–97.

39. Dempster AP, Laird NM, Rubin DB. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society. 1977;Series B(39):1–38.

40. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999;401(6755):788–91.

41. Pauca VP, Piper J, Plemmons RJ. Nonnegative matrix factorization for spectral data analysis. Linear Algebra and Its Applications. 2006;416(1):29–47.

42. Aggarwal CC. Recommender systems: the textbook. Switzerland: Springer International Publishing; 2016. 1–493 p.

43. Lin XE, Boutros P. NNLM: a package for fast and versatile nonnegative matrix factorization. 2019. Available from: https://github.com/linxihui/NNLM

44. Chen R, Varshney L. Non-negative matrix factorization of clustered data with missing values. In: IEEE Data Science Workshop (DSW). 2019. p. 180–4.

45. Jiang S, Wang C, Weiss DJ. Sample size requirements for estimation of item parameters in the multidimensional graded response model. Frontiers in Psychology. 2016;7(February).

46. Olsbjerg M, Christensen KB. Modeling local dependence in longitudinal IRT models. Behavior Research Methods. 2015;47(4):1413–24.

47. De Ayala RJ. The influence of multidimensionality on the graded response model. Applied Psychological Measurement. 1994;18(2):155–70.

48. Bulut O, Sunbul Ö. Monte Carlo Simulation Studies in Item Response Theory with the R Programming Language. Journal of Measurement and Evaluation in Education and Psychology. 2017;8(3):266–87.

49. Finch HW. The impact of missing data on the detection of nonuniform differential item functioning. Educational and Psychological Measurement. 2011;71(4):663–83.

50. Schouten RM, Lugtig P, Vink G. Generating missing values for simulation purposes: a multivariate amputation procedure. Journal of Statistical Computation and Simulation. 2018;88(15):2909–30.

51. Nassiri V, Molenberghs G, Verbeke G, Barbosa-Breda J. Iterative multiple mmputation: a framework to determine the number of imputed datasets. American Statistician. 2020;74(2):125–36.

52. Goretzko D. Factor retention in exploratory factor analysis with missing data. Educational and Psychological Measurement. 2021;1–21.

53. van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. Journal of Statistical Software. 2011;45(3):1–67.

54. Chen PY, Wu W, Garnier-Villarreal M, Kite BA, Jia F. Testing measurement invariance with ordinal missing data: a comparison of estimators and missing data techniques. Multivariate Behavioral Research. 2020;55(1):87–101.

55. Chalmers RP. mirt: a multidimensional item response theory package for the R environment. Journal of Statistical Software. 2012;48(6):1–29.

56. Bradley J V. Robustness. British Journal of Mathematical & Statistical Psychology. 1978;31(2):144–52.

57. Kaplan D. A study of the sampling variability and z-values of parameter estimates from misspecified structural equation models. Multivariate Behavioral Research. 1989;24(1):41–57.

58. Curran P, West SG. The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. Psychological Methods. 1996;1(1):16–29.

59. Kopec JA, Sayre EC, Davis AM, Badley EM, Abrahamowicz M, Sherlock L, et al. Assessment of health-related quality of life in arthritis: conceptualization and development of five item banks using item response theory. Health and Quality of Life Outcomes. 2006;4.

60. Kopec JA, Badii M, Mckenna M, Lima VD, Sayre EC, Dvorak M. Computerized adaptive testing in back pain validation of the CAT-5D-QOL. Vol. 33, SPINE. 2008.

61. Sawatzky R, Ratner P, Kopec J, Wu A, Zumbo B. The accuracy of computerized adaptive testing in heterogeneous populations: A mixture item-response theory analysis. PLoS One. 2016;11(3).

62. Sindhu BS, Wang YC, Lehman LA, Hart DL. Differential item functioning in a computerized adaptive test of functional status for people with shoulder impairments is negligible across pain intensity, gender, and age groups. OTJR: Occupation, Participation and Health. 2013;33(2):86–99.

63. Teresi JA, Fleishman JA. Differential item functioning and health assessment. Quality of Life Research. 2007;16(SUPPL. 1):33–42.

64. Slocum-Gori SL, Zumbo BD. Assessing the unidimensionality of psychological scales: using multiple criteria from factor analysis. Social Indicators Research. 2011;102(3):443–61.

65. Reise SP, Cook KF, Moore TM. Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In: Reise SP, Revicki. DA, editors. Handbook of item response theory modeling: applications to typical performance assessment. New York: Routledge; 2014. p. 13–37.

66. Browne MW, Cudeck R. Alternative ways of assessing model fit. Sociological Methods & Research. 1992;21(2):230–58.

67. Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Structural Equation Modeling. 1999;6(1):1–55.

68. Hochberg Yosef. A sharper Bonferroni procedure for multiple tests of significance. Biometrika. 1988;75(4):800–2.

69. Meade AW, Wright NA. Solving the measurement invariance anchor item problem in item response theory. Journal of Applied Psychology. 2012;97(5):1016–31.

70. Kopf J, Zeileis A, Strobl C. Anchor selection strategies for DIF analysis: review, assessment, and new approaches. Vol. 75, Educational and Psychological Measurement. 2015. 22–56 p.

71. Sedivy SK, Zhang B, Traxel NM. Detection of differential item functioning with polytomous items in the presence of missing data. In: Annual meeting of the National Council on Measurement in Education. 2006.

72. Finch HW. The impact of missing data on the detection of nonuniform differential item functioning. Educational and Psychological Measurement. 2011;71(4):663–83.

73. Lopez Rivas GE, Stark S, Chernyshenko OS. The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test. Applied Psychological Measurement. 2009;33(4):251–65.

Supplementary Materials

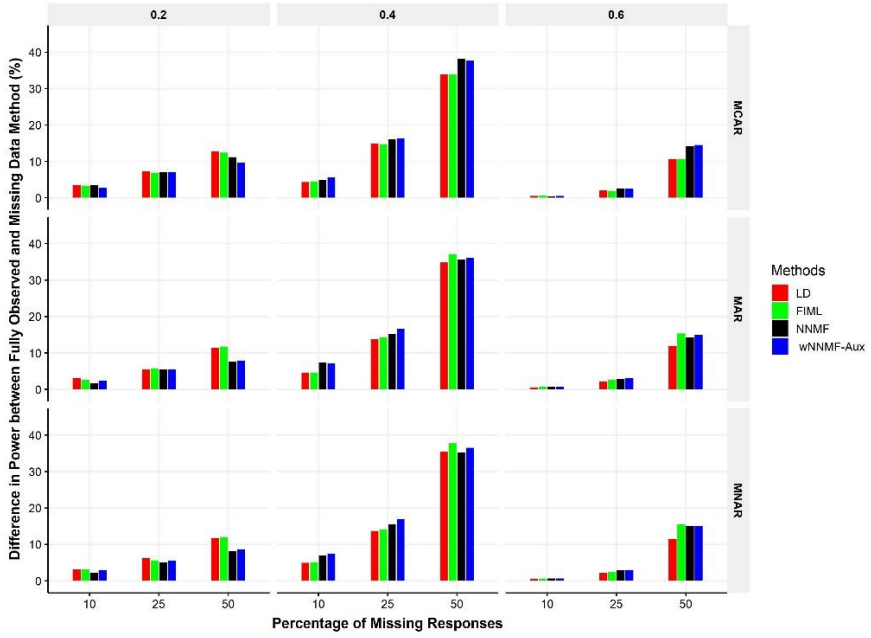
Enhanced weighted non-negative matrix factorization approach for addressing item non-response in patient-reported outcome measures

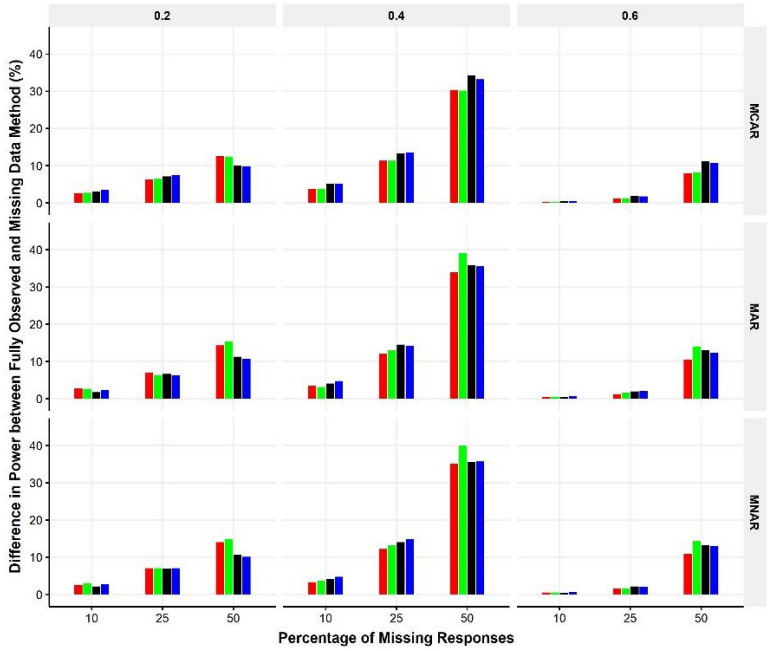
Supplementary Material 1: Statistical analyses of the CAT-5D-QOL PAIN domain

The all-other anchor approach (AOAA), which uses all items except for the currently studied item as the anchor, was used to select one anchor item in order to define the latent construct on which the groups are compared (73). First, we fit a model (i.e., the model with the best fit statistics) to the data in which the slopes and thresholds of all the items were constrained equal across the two groups (i.e., males/females). Second, a series of unconstrained models were fitted to the data. The LRT was used to test the difference between the constrained and unconstrained models for each item. We labelled an item as DIF-free if the LRT statistic was not statistically significant. Next, we selected the DIF-free items that had the largest slope parameters as the anchor items for the dimension. This approach has been shown to outperform other anchor selection strategies (69,70).

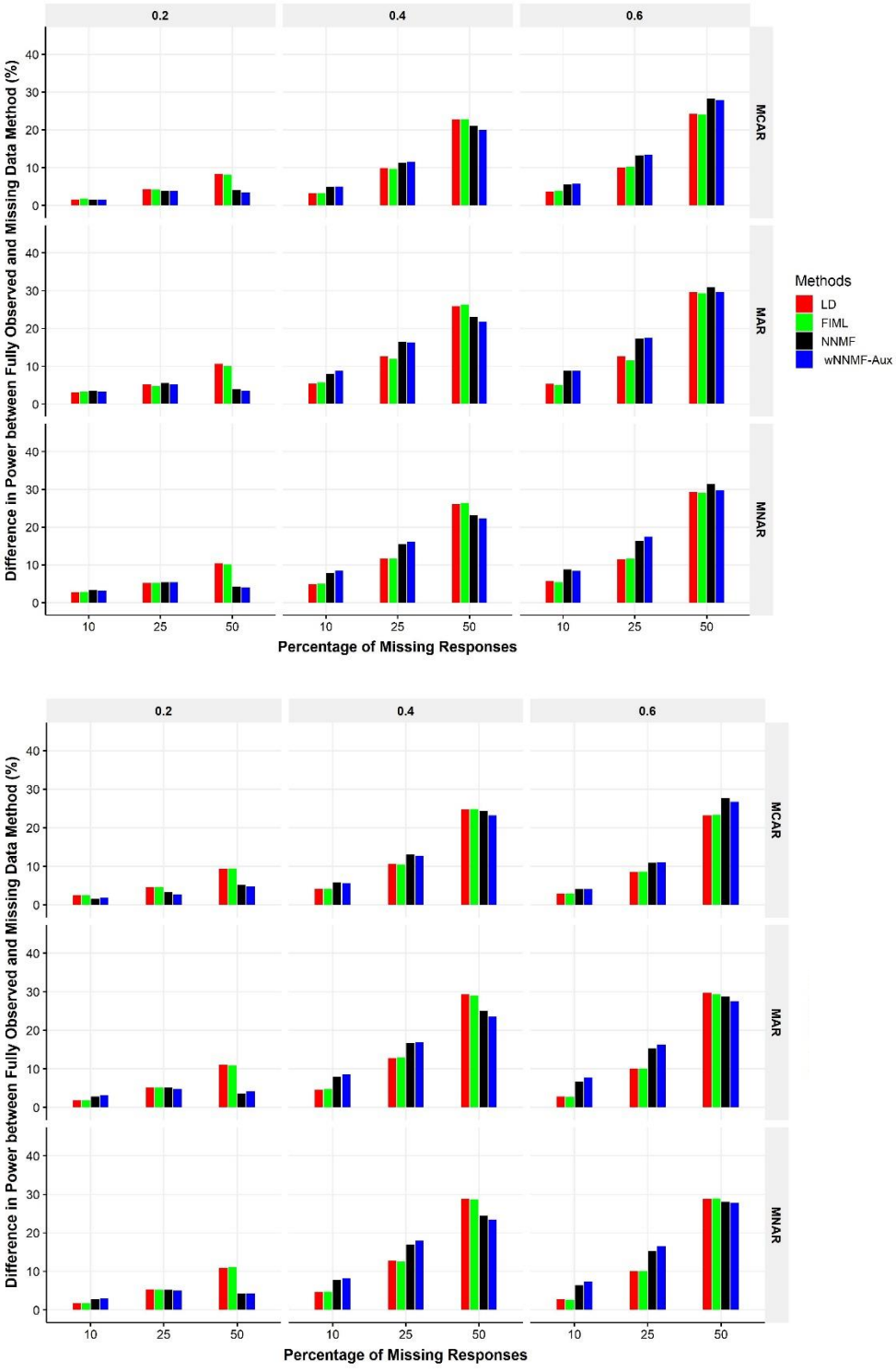
For the assessment of uniform and non-uniform DIF of the items on the PAIN domain, we specified an unconstrained DIF model, where the slopes and thresholds of all items, except for the anchor item, were freely estimated. Next, we fitted a set of constrained models (i.e., no-DIF models), where the parameters of the remaining items were constrained one at a time. Similarly, the LRT statistic was used to test the difference between the unconstrained DIF and no-DIF models. A statistically significant LRT statistic when an unconstrained DIF model was compared to: (1) a no-DIF model (i.e., the threshold parameters were constrained for the two groups), indicates the presence of uniform DIF, and (2) a no-DIF model, where the slope parameters are constrained for the two groups, indicates the presence of non-uniform DIF.

Supplementary Material 2: Difference in Power of the LRT to detect uniform and non-uniform DIF

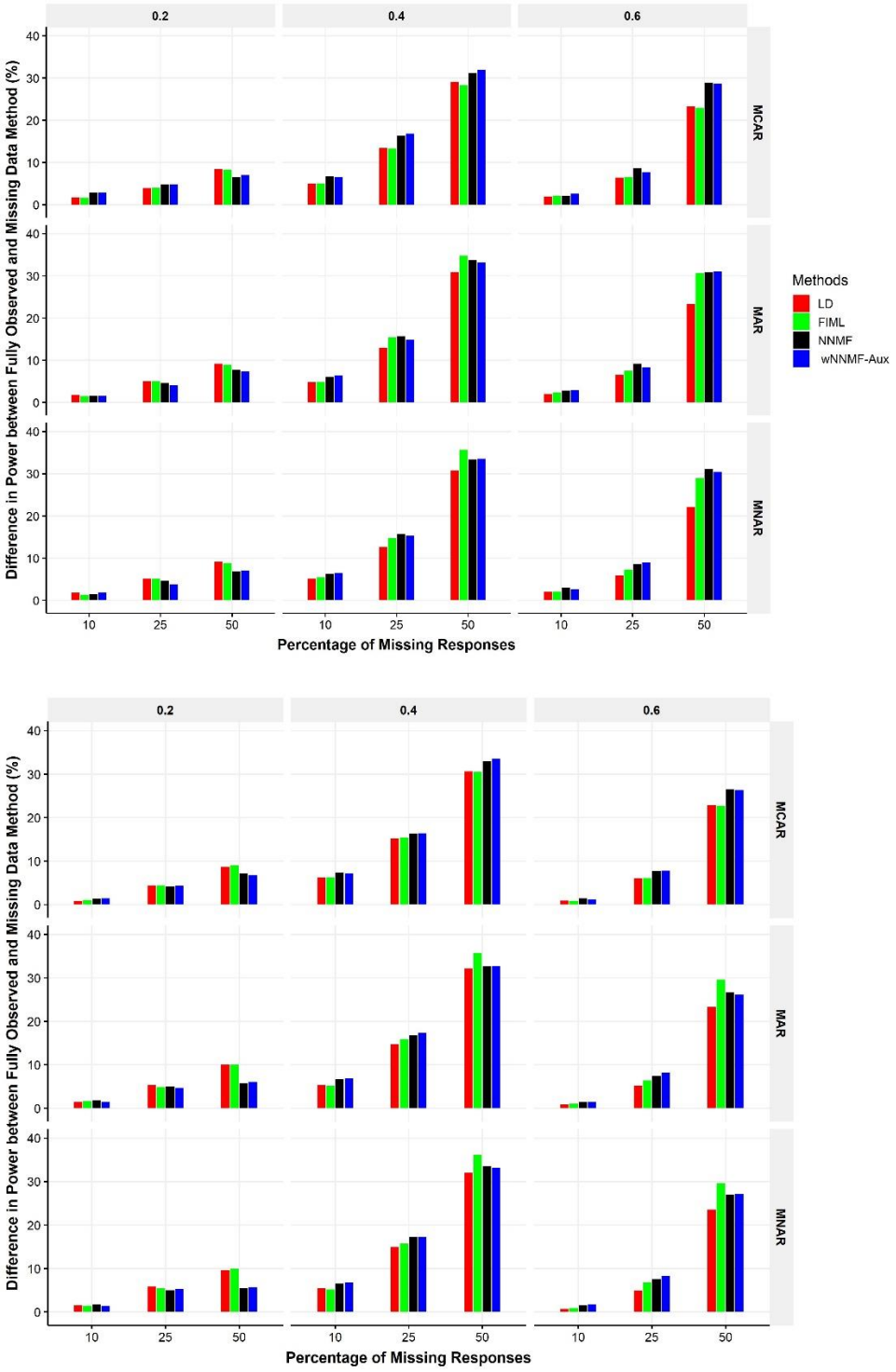




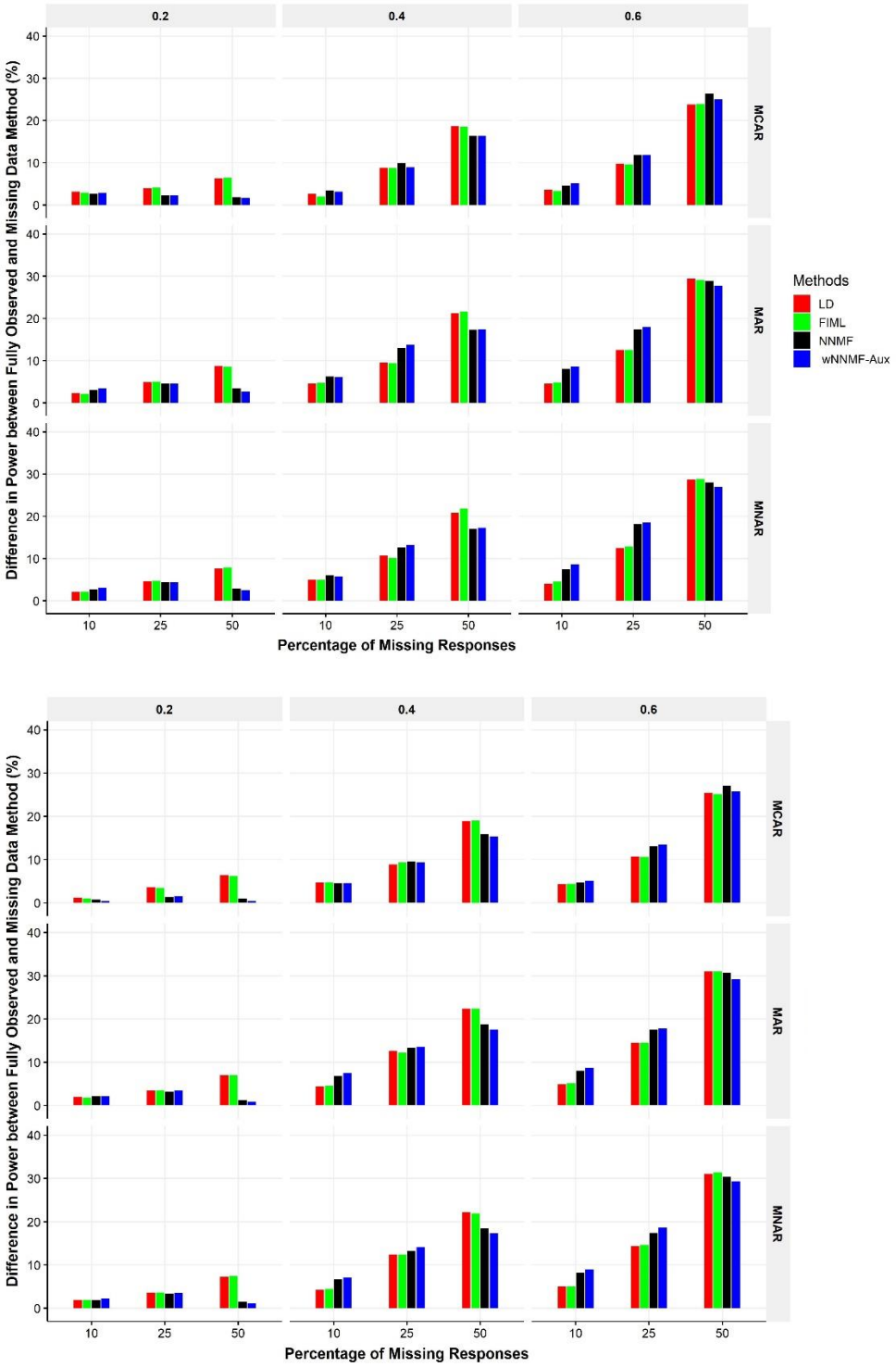
Supplementary Figure 6-1. Difference in Power of the LRT to detect **uniform DIF** between fully observed and missing data method (%) for 15 (top panel) and 30 items (bottom panel), and sample size of R1500/F500 (R: reference group; F: focal group) stratified by the missing data mechanism and magnitude of DIF. MCAR: missing completely at random; MAR: missing at random; and MNAR: missing not at random.



Supplementary Figure 6-2. Difference in Power of the LRT to detect **non-uniform DIF** between fully observed and missing data method (%) for 15 (top panel) and 30 items (bottom panel), and sample size of R1500/F500 (R: reference group; F: focal group) stratified by the missing data mechanism and magnitude of DIF. MCAR: missing completely at random; MAR: missing at random; and MNAR: missing not at random.



Supplementary Figure 6-3. Difference in Power of the LRT to detect **uniform DIF** between fully observed and missing data method (%) for 15 (top panel) and 30 items (bottom panel), and sample size of R500/F500 (R: reference group; F: focal group) stratified by the missing data mechanism and magnitude of DIF. MCAR: missing completely at random; MAR: missing at random; and MNAR: missing not at random.



Supplementary Figure 6-4. Difference in Power of the LRT to detect **non-uniform DIF** between fully observed and missing data method (%) for 15 (top panel) and 30 items (bottom panel), and sample size of R500/F500 (R: reference group; F: focal group) stratified by the missing data mechanism and magnitude of DIF. MCAR: missing completely at random; MAR: missing at random; and MNAR: missing not at random.

Supplementary Material 3: Distribution of item responses

Supplementary Table 1: Distribution of item responses (%) on the CAT-5D-QOL pain item bank (*N* = 1660)

Item	Response Options					
	Not at all	**A little bit**	**Moderately**	**Quite a bit**	**Extremely**	**Missing**
Q141: Pain interference with normal work	618 (37.1)	469 (28.2)	218 (13.1)	222 (13.3)	116 (7.0)	17 (1.0)
Q142: Pain affect ability to fall asleep	703 (42.2)	452 (27.1)	229 (13.7)	204 (12.2)	57 (3.4)	15 (0.9)
Q143: Pain interference with physical leisure	576 (34.6)	415 (24.9)	199 (11.9)	230 (13.8)	215 (12.9)	25 (1.5)
Q146: Pain interference with non-physical leisure	997 (59.8)	394 (23.6)	157 (9.4)	78 (4.7)	12 (0.7)	22 (1.3)
Q147: Pain interference with social activities	937 (56.2)	373 (22.4)	180 (10.8)	116 (7.0)	32 (1.9)	28 (1.3)
Q149: Pain interference with self-care activities	1167 (70.0)	300 (18.0)	96 (5.8)	63 (3.8)	15 (0.9)	19 (1.1)
Q150: Pain affect ability to use the toilet	1361 (81.7)	196 (11.8)	48 (2.9)	33 (2.0)	22 (1.3)	
Q151: Pain interference with light household chores	1050 (63.0)	335 (20.1)	138 (8.3)	84 (5.0)	26 (1.6)	33 (1.6)
Q154: Pain interference with physical leisure	507 (30.4)	316 (19.0)	137 (8.2)	219 (13.1)	436 (26.2)	45 (2.7)
Q157: Pain interference with sleep	630 (37.8)	501 (30.1)	260 (15.6)	200 (12.0)	53 (3.2)	16 (1.0)
Q158: Pain interference with social activities outside the home	840 (50.4)	373 (22.4)	197 (11.8)	166 (10.0)	62 (3.7)	22 (1.3)
Q159: Pain interference with eating meals	1309 (78.6)	221 (13.3)	81 (4.9)	32 (1.9)	17 (1.0)	
Q160: Pain affect ability to accomplish more	656 (39.4)	438 (26.3)	188 (11.3)	219 (13.1)	125 (7.5)	34 (2.0)
Q169: Pain limitation in doing things	600 (36.0)	422 (25.3)	230 (13.8)	258 (15.5)	133 (8.0)	17 (1.0)
Q170: Pain frustration	453 (27.2)	460 (27.6)	258 (15.5)	297 (17.8)	174 (10.4)	18 (1.1)
Q173: Pain restrict from routine daily activities	713 (42.8)	410 (24.6)	217 (13.0)	232 (13.9)	68 (4.1)	20 (1.2)
	Never	**Seldom**	**Sometimes**	**Often**	**Always**	**Missing**
Q144: Intense pain	636 (38.2)	315 (18.9)	358 (21.5)	259 (15.5)	73 (4.4)	19 (1.1)
Q148: Pain prevent from taking bath or shower	1276 (76.6)	175 (10.5)	121 (7.3)	52 (3.1)	13 (0.8)	23 (1.4)
Q152: Pain prevent from dressing and undressing	1326 (79.6)	171 (10.3)	99 (5.9)	35 (2.1)	13 (0.8)	16 (1.0)

Q153: Pain prevent from using the toilet	1520 (91.2)	79 (4.7)	46 (2.8)			15 (0.9)
Q155: Pain prevent from grooming	1379 (82.8)	148 (8.9)	76 (4.6)	39 (2.3)		18 (1.1)
Q166: Pain prevent from eating meal	1357 (81.5)	164 (9.8)	102 (6.1)	23 (1.4)		14 (0.8)
Q167: Pain restrict from routine daily activities	742 (44.5)	317 (19.0)	333 (20.0)	181 (10.9)	70 (4.2)	17 (1.0)
Q171: Pain prevent from routine daily activities	688 (41.3)	333 (20.0)	383 (23.0)	203 (12.2)	37 (2.2)	16 (1.0)
Q172: Wake up at night due to pain	579 (34.8)	351 (21.1)	414 (24.8)	227 (13.6)	73 (4.4)	16 (1.0)
Q175: Help in dressing or bathing due to pain	1397 (83.9)	114 (6.8)	75 (4.5)	31 (1.9)	22 (1.3)	21 (1.3)
Q176: Stay in bed because of pain	1207 (72.4)	207 (12.4)	155 (9.3)	62 (3.7)		29 (1.7)
Q177: Unbearable pain	1052 (63.1)	232 (13.9)	226 (13.6)	120 (7.2)		30 (1.8)
	No pain	**Mild pain**	**Moderate pain**	**Severe pain**		**Missing**
Q138: Select one that describes you best	298 (17.9)	669 (40.2)	452 (27.1)	218 (13.1)		23 (1.4)
	None	**Very mild**	**Mild**	**Moderate**	**Severe**	**Missing**
Q139: How much bodily pain have you had?	237 (14.2)	414 (24.8)	285 (17.1)	432 (25.9)	275 (16.5)	17 (1.0)
	None	**A few**	**Some**	**Most**		**Missing**
Q140: Pain interference in a number of activities	725 (43.5)	401 (24.1)	309 (18.5)	204 (12.2)		21 (1.3)
	Free of pain	**Mild-moderate pain**	**Moderate pain**	**Moderate-severe pain**	**Severe pain**	**Missing**
Q174: Which one best describes level of pain?	283 (17.0)	524 (31.5)	402 (24.1)	328 (19.7)	98 (5.9)	25 (1.5)
	No pain	**Moderate pain**	**Extreme pain**			**Missing**
Q178: Which statement best describes your pain?	400 (24.0)	1060 (63.6)	154 (9.2)			46 (2.8)
	None of the time	**A little of the time**	**Some of the time**	**Most of the time**	**All of the time**	**Missing**
Q164: Feel great physically	384 (23.0)	253 (15.2)	263 (15.8)	607 (36.4)	136 (8.2)	17 (1.0)
Q168: Avoid strenuous activities for fear of pain	602 (36.1)	287 (17.2)	246 (14.8)	287 (17.2)	217 (13.0)	21 (1.3)

CHAPTER 7: SUMMARY

7.1 Summary of Research Findings

The purpose of this thesis was to develop and evaluate methods to address item non-response in patient-reported outcome measures (PROMs). The objectives were to: (1) compare the performance of ordinal item non-response methods when testing for differential item functioning (DIF); (2) evaluate the performance of ordinal item non-response methods when estimating change in latent variable means over time; (3) estimate longitudinal change in PROMs scores in the presence of missing data using true-score model; and (4) develop a new method to address ordinal item non-response and compare its performance with several ordinal item non-response methods. The objectives can be grouped into two parts.

In the first part, the non-negative matrix factorization (NNMF) method, an empirical method based on dependencies that exist among ordinal items to address item non-response without making any distributional assumption, was investigated [1, 2]. This method was explored in both cross-sectional and longitudinal data.

In the cross-sectional data, Type I error rates and statistical power to detect DIF effects in multidimensional PROMs for the NNMF method were compared to the full information maximum likelihood (FIML), listwise deletion (LD) and half-mean imputation (HMI) methods using simulated and real-world data. Monte Carlo simulation conditions included sample size, percentage of missing data, magnitude of uniform and non-uniform DIF effects, and magnitude of correlation between latent (i.e., unobserved) variables. The performance of the NNMF method was similar to the FIML and LD methods when sample size was large (i.e., $N > 500$). However,

when the sample size was relatively small (i.e., $N = 500$) and the percentage of missing responses was high (i.e., 50%), the NNMF method resulted in inflated error rates.

The Type I error rates for the NNMF, LD and FIML methods were controlled within the bounds of Bradley's liberal criterion (i.e., 0.025 – 0.075) for $\alpha = 0.05$ [3] for most of the simulation conditions under consideration. Statistical power to detect non-uniform DIF increased as the latent variable correlation increased; this trend was not observed for uniform DIF. Overall, the power to detect uniform DIF was greater than the power to detect non-uniform DIF. These findings on the detection of uniform and non-uniform in the presence of item non-response are consistent with previous studies [4–6].

For longitudinal data, Monte Carlo simulations and a numeric example were used to evaluate the bias and mean squared error (MSE) of the NNMF, multiple imputation with conditional proportional odds model (POM), FIML, and LD methods when estimating change in latent variable means between two measurement occasions. This study was conducted in the context of longitudinal latent variable measurement models under the assumption of longitudinal measurement invariance. Simulation conditions included sample size, percentage of missing data, and magnitude of the correlation between the first and second measurement occasions. A population-based clinical-registry of patients having total-hip-arthroplasty was used as a numeric example.

The NNMF method outperformed the FIML and POM methods when the correlation between the first and second measurement occasions was strong (i.e., $\rho = 0.7$). The MSE of the NNMF method reduced significantly as sample size increased when compared with other methods. The NNMF method did not perform as well as the FIML and POM methods when the correlation between the first and second measurement occasions was weak (i.e., $\rho = 0.3$). Consistent with

previous research findings, the performance of the FIML and POM methods were similar for all simulation conditions [7, 8]. In the numeric example, the direction of estimated change in latent variable mean over time was similar for all the item non-response methods under consideration. The NNMF method resulted in a smaller estimated mean change and standard error when compared with other item non-response methods.

The second part of this thesis investigated the use of auxiliary variable in missing data methods, and an enhanced weighted NNMF (wNNMF-Aux) method, which uses aggregated observed item responses as auxiliary variable to define weights for item level imputation, was proposed. The impact of auxiliary variable on the performance of different missing data methods were examined at the summary scores and item levels of PROMs.

At the summary scores level, the impact of auxiliary variables on the bias and precision of the multiple imputation (MI) was investigated and compared with the LD and FIML methods when estimating longitudinal change in PROMs scores using true-score model. A single hypothetical auxiliary variable was generated to correlate with both pre- and post-operative PROMs scores. The study was conducted using a population-based joint replacement clinical registry and simulated data.

Precision and bias of the model parameters were estimated under varying sample sizes, magnitude of correlation, and percentage of missing data in the simulation study. The precision of the estimated parameters increased significantly when the correlation between the auxiliary variable and the outcome of interest was 0.5 or more and the percentage of missing data was 25% or more [7, 9, 10]. Also, including an auxiliary variable in the imputation model moderated the bias and size of the root mean squared error when data were missing not at random.

At the item level, the Type I error rates and statistical power to detect DIF effects in unidimensional (i.e., one latent construct or variable) PROMs for the wNNMF-Aux method were compared to the FIML, NNMF and LD methods using Monte Carlo simulation and real-world data. The simulation conditions included number of items, sample size, magnitude of DIF, and percentage of missing observations. The computerized adaptive test (CAT-5D-QOL) was used as a real-world example to demonstrate the implementation of the item non-response methods. The sum of the observed item responses was used to define an auxiliary variable for the wNNMF-Aux method [11, 12].

The wNNMF-Aux method outperformed the NNMF method for most of the simulation conditions under consideration, especially when the number of items was large (i.e., 30) and the percentage of missing responses was high (i.e., 50%). The performance of the wNNMF-Aux method was similar to the FIML and LD methods for large number of items. However, when the number of items was small (i.e., 15) and the percentage of missing responses was high, the wNNMF-Aux method resulted in inflated error rates. The performance of all the item non-response methods considered were similar in the analysis of the CAT-5D-QOL data.

Table 7 provides a summary of the various methods considered across study objectives and simulation conditions. All the missing data methods considered in this research can be implemented in RStudio. However, the wNNMF and wNNMF-Aux methods are not readily available in RStudio.

Table 7-1. Summary of performance of missing data methods across study objectives

Study Objective	Methods	Sample size (*N*)	% of missing responses	Latent variable correlation	Aux. variable correlation with PROMs scores	Performance	Software
#1: To compare the performance of ordinal item non-response methods when testing for DIF	NNMF FIML LD HMI	500 to 2000	10%, 25% and 50%	ρ = 0.3 and 0.5	Not applicable	i. NNMF compared (Type I error rates; power rates) favorably to the LD and FIML methods when sample size was large (i.e., $N > 500$) ii. NNMF resulted in inflated Type I error rates with relatively small sample size (i.e., $N \leq 500$)	All methods can be easily implemented in RStudio
#2: To evaluate the performance ordinal item non-response methods when estimating latent variable means over time	NNMF FIML POM LD	500 and 1000	10%, 25% and 50%	ρ = 0.3 and 0.7	Not applicable	i. The best performance (absolute percentage bias, mean squared error, and relative efficiency) of the NNMF method was observed for conditions with large sample size (i.e., N > 500), high latent variable correlation (i.e., ρ = 0.7), and percentage of item non-response ≥ 25% ii. Performance of the LD method were consistently poor across simulation conditions	All methods can be easily implemented in RStudio
#3: To estimate change in PROMs scores over time in	FIML MI MI-Aux	1000	10%, 25% and 50%	Not applicable	ρ = 0.2, 0.5 and 0.8	i. Including an auxiliary variable that is strongly correlated (i.e., $\rho \geq 0.5$)	All methods can be easily

the presence of missing data using true-score model	LD						with the outcome of interest in the imputation phase of the MI method reduced the bias and error of the estimated parameters ii. Including an auxiliary variable that is strongly correlated (i.e., $\rho \geq 0.5$) with the outcome of interest in the imputation phase of the MI method increased the precision of the estimated parameters	implemented in RStudio
#4: To develop a new method to address ordinal item non-response and compare its performance with several ordinal item non-response methods	wNNMF-Aux NNMF FIML LD	1000 to 2000	10%, 25% and 50%	Not applicable	Not applicable	i. ii.	The wNNMF-Aux method showed comparable performance (Type I error rates and Power) to the LD and FIML methods across simulation conditions The wNNMF-Aux outperformed NNMF for most of the simulation conditions, especially when the number of items was large (say, 30 items)	All methods can be implemented in RStudio. However, wNNMF and wNNMF-Aux methods are not readily available in RStudio

LD: List-wise deletion; HMI: Half-mean imputation; MI: Multiple Imputation; MI-Aux: Multiple Imputation with Auxiliary Variable; POM: fully conditional specification proportional odds model; FIML: full information maximum likelihood; NNMF: Non-negative matrix factorization; wNNMF-Aux: Enhanced Weighted Non-negative matrix factorization; ρ: correlation between latent variables or time-specific latent variable correlation or auxiliary variable correlation with PROMs scores; N: total sample size

7.2 Strengths and Limitations

This research is a novel addition to the literature on statistical methods for PROMs via its focus on proposing and evaluating distribution-free and computational tractable unsupervised machine-learning methods, namely the NNMF and wNNMF-Aux methods, to address ordinal item non-response in PROMs. Existing methods to address ordinal item non-response including FIML and MI are based on a family of parametric probability model assumptions, which are often violated when item responses are ordinal [13, 14]. Also, NNMF and wNNMF-Aux captures any potential non-linear relationship that may exist amongst items when imputing missing responses. Investigating psychometric properties at the item level can provide an understanding of the impact of each item to the subscale or scale.

This research showed the practicality of using existing data to define an auxiliary variable, which provides a basis for easy, accessible, and cost-effective approach of identifying auxiliary variable. Leveraging on the information from the auxiliary variable ensures that the missing at random mechanism assumption is realistic. When the missing data mechanism suggests that data are missing not at random, incorporating information from auxiliary variables into the missing data methods can help reduce bias and improve statistical power.

The use of both simulated and real-world data to investigate the performance of several item non-response methods further strengthens this research. A wide variety of data conditions including sample sizes, percentage of missing responses and missingness mechanisms were investigated through the simulation study. Practical examples were illustrated using population-based joint replacement clinical registry from Manitoba [15, 16], and a heterogenous cohort, which includes rheumatology clinic, the joint replacement waiting list, and stratified-random community samples from British Columbia [17].

There are, however, some limitations in this research. First, the NNMF method performs poorly when the correlations amongst items are weak. This is because the NNMF method relies on the associations amongst items to predict missing responses. Weak correlations imply the associations amongst items are not sufficiently strong to accurately predict missing responses [18–20]. Second, the low rank of the item response matrix needs to be specified. The low rank was specified in this research because the number of dimensions expected in the PROMs instruments under consideration were known *a priori*. A data-driven method or exploratory factor analysis could be used to identify the low rank when there is no prior information about the number of dimensions [21]. Also, the variables considered to be associated with DIF in this research were known *a priori*. It is possible that the study population are heterogenous with respect to an unobserved characteristic. Hence, latent variable mixture models could be used to identify homogenous sub-groups with the population [22, 23]. Third, like other studies, the simulation conditions investigated were not exhaustive. The number of item response categories, number of items, and total sample size were limited; larger numbers for these conditions would substantially increase computational time. The simulations conditions in this research were consistent with previous studies [6, 7, 24, 25].

7.3 Future Directions

This research has investigated several ordinal item non-response methods in PROMs and proposed an enhanced weighted NNMF method. There are, however, several topic areas for future considerations. These include: (1) establishing the relationship between ordinal measured and latent variables in PROMs using the NNMF method, and (2) combining the NNMF method with item response theory (IRT) models to test for DIF, and (3) developing R package for the wNNMF and wNNMF-Aux methods.

Establishing the association between ordinal measured and latent variables is of interest in PROMs. While several studies have used exploratory and confirmatory factor analyses, with maximum likelihood estimation, which assumes a multivariate normal distribution, to establish the relationship between measured and latent variables association [26–29], future research could investigate non-parametric methods such as the NNMF method. However, obtaining a unique solution using the NNMF method can be challenging.

DIF, which occurs when individuals with the same underlying level of health have different measured item response probabilities, can lead to lack of scale comparability and erroneous conclusions about the presence of group differences, if unaccounted for in the analysis of PROMs. A number of studies have used the combination of IRT models and model-based recursive partitioning techniques (IRT-RPTs) to test for DIF [30–32]. The detection of DIF using the IRT-RPTs is data-driven and it appears promising. However, it is computationally intensive because it is a tree-based method. Future research could combine a computationally efficient method, such as the NNMF method, with IRT models to test for DIF while taking into account item non-response.

Availability and accessibility of computer programs or codes to implement statistical methods can increase the visibility of research article. A number of R packages including *NNLM* [33] and *NMF* [34] are available to implement the NNMF method, but not wNNMF and wNNMF-Aux methods. Future work could consider developing an R package for the wNNMF and wNNMF-Aux methods in order to make these methods readily available for researchers.

7.4 Conclusions and Recommendations

This research investigated several methods to address ordinal item non-response in PROMs and identify the NNMF method as a promising method to address item level non-response because of

it is non-parametric and computationally efficient. The NNMF method has been applied in other research areas such as recommendation systems, image and signal processing [18, 35–38]; this is the first study to use the NNMF method to address ordinal item non-response in PROMs, and propose an enhanced wNNMF-Aux method, which incorporates auxiliary information in defining its weight.

The NNMF and wNNMF-Aux methods compared favorably to the LD and FIML methods when testing for DIF. However, the Type I error rates were inflated as sample size decreased. When estimating longitudinal change in latent variable means, the best performance of the NNMF method was observed in the large sample size condition ($N > 500$) and when the percentage of item non-response was 25% or more. When estimating longitudinal change in PROMs scores using true-score model, including an auxiliary variable with strong correlation with the outcome of interest in the imputation phase of the MI method increased the precision of the estimated parameters.

Overall, there was no single method that performed best in all the simulation scenarios. Hence, researchers should consider comparing two or more methods in sensitivity analyses. Also, under the expectation of inevitable missing observations in PROMs data, researchers should plan to include auxiliary information that may be associated with the outcome(s) of interest in their study.

This research contributes to the statistical literature on methods to address missing data with potential applications in clinical and quality of life research. Also, it provides guidance on factors to consider when selecting methods to address missing data in PROMs. Finally, findings from this research would be beneficial to psychometricians and other researchers who use PROMs to

understand the complex relationship that exist among ordinal items and underlying latent variables.

7.5 References

1. Aggarwal, C., & Parthasarathy, S. (2001). Mining massively incomplete data sets by conceptual reconstruction. In *ACM KDD* (pp. 227–232).

2. Agarwal, D., & Chen, B. C. (2009). Regression-based latent factor models. In *ACM KDD International Conference on Knowledge Discovery and Data Mining* (pp. 19–28).

3. Bradley, J. V. (1978). Robustness. *British Journal of Mathematical & Statistical Psychology*, *31*(2), 144–152.

4. Sedivy, S. K., Zhang, B., & Traxel, N. M. (2006). Detection of differential item functioning with polytomous items in the presence of missing data. In *Annual meeting of the National Council on Measurement in Education*.

5. Finch, H. W. (2011). The impact of missing data on the detection of nonuniform differential item functioning. *Educational and Psychological Measurement*, *71*(4), 663–683.

6. Finch, H. (2011). The use of multiple imputation for missing data in uniform DIF analysis: power and type I error rates. *Applied Measurement in Education*, *24*(4), 281–301.

7. Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351.

8. Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177.

9. Bell, M. L., & Fairclough, D. L. (2014). Practical and statistical issues in missing data for longitudinal patient-reported outcomes. *Statistical Methods in Medical Research*, *23*(5), 440–459.

10. Wang, C., & Hall, C. B. (2010). Correction of bias from non-random missing longitudinal data using auxiliary information. *Stat Med*, *29*(6), 671–679.

11. Eekhout, I., Enders, C. K., Twisk, J. W. R., de Boer, M. R., de Vet, H. C. W., & Heymans, M. W. (2015). Analyzing incomplete item scores in longitudinal data by including item score information as auxiliary variables. *Structural Equation Modeling*, *22*(4), 588–602.

12. Enders, C. K. (2008). A note on the use of missing auxiliary variables in full information maximum likelihood-based structural equation models. *Structural Equation Modeling*, *15*(3), 434–448.

13. Donneau, A. F., Mauer, M., Molenberghs, G., & Albert, A. (2015). A simulation study comparing multiple imputation methods for incomplete longitudinal ordinal data. *Communications in Statistics: Simulation and Computation*, *44*(5), 1311–1338.

14. Donneau, A. F., Mauer, M., Lambert, P., Molenberghs, G., & Albert, A. (2015). Simulation-based study comparing multiple imputation methods for non-monotone missing ordinal data in longitudinal settings. *Journal of Biopharmaceutical Statistics, 25*(3), 570–601.

15. Yadegari, I., Bohm, E., Ayilara, O. F., Zhang, L., Sawatzky, R., Sajobi, T. T., & Lix, L. M. (2019). Differential item functioning of the SF-12 in a population-based regional joint replacement registry. *Health and Quality of Life Outcomes*, *17*(1), 1–11.

16. Zhang, L., Lix, L. M., Ayilara, O., Sawatzky, R., & Bohm, E. R. (2018). The effect of multimorbidity on changes in health-related quality of life following hip and knee arthroplasty. *Bone and Joint Journal*, *100B*(9), 1168–1174.

17. Sawatzky, R., Ratner, P., Kopec, J., Wu, A., & Zumbo, B. (2016). The accuracy of computerized adaptive testing in heterogeneous populations: a mixture item-response theory analysis. *PloS one*, *11*(3).

18. Zhang, S., Wang, W., Ford, J., & Makedon, F. (2006). Learning from incomplete ratings using non-negative matrix factorization. In *Proceedings of the Sixth SIAM International Conference on Data Mining* (pp. 549–553).

19. Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791.

20. Mazumder, R., Hastie, T., & Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, *11*, 2287–2322.

21. Lin, X. E., & Boutros, P. C. (2020). Optimization and expansion of non-negative matrix factorization. *BMC Bioinformatics*, *21*(1), 1–10.

22. Sawatzky, R., Ratner, P. A., Kopec, J. A., & Zumbo, B. D. (2012). Latent variable mixture models: a promising approach for the validation of patient reported outcomes. *Quality of Life Research*, *21*(4), 637–650.

23. Lix, L. M., & Ayilara, O. (2022). Latent variable mixture models to address heterogeneity in patient-reported outcome data. *Methods*, *204*, 151–159.

24. Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: a comparison of estimators. *Structural Equation Modeling*, *21*(2), 167–180.

25. Bulut, O., & Sunbul, Ö. (2017). Monte Carlo simulation studies in item response theory with the R programming language. *Journal of Measurement and Evaluation in Education and Psychology*, *8*(3), 266–287.

26. Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299.

27. Conway, J. M., & Huffcutt, A. I. (2003). A review and evaluation of exploratory factor analysis practices in organizational research. *Organizational Research Methods*, *6*(2), 147–168.

28. Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: common errors and some comment on improved practice. *Educational and Psychological Measurement*, *66*(3), 393–416.

29. Fernandez, A. C., Fehon, D. C., Treloar, H., Ng, R., & Sledge, W. H. (2015). Resilience in organ transplantation: an application of the Connor-Davidson resilience scale (CD-RISC) with liver transplant candidates. *J Pers Assess*, *97*(5), 487–493.

30. Henninger, M., Debelak, R., & Strobl, C. (2022). A new stopping criterion for Rasch trees based on the Mantel–Haenszel effect size measure for differential item functioning. *Educational and Psychological Measurement*.

31. Wickelmaier, F., & Zeileis, A. (2018). Using recursive partitioning to account for parameter heterogeneity in multinomial processing tree models. *Behavior Research Methods*, *50*(3), 1217–1233.

32. Komboz, B., Strobl, C., & Zeileis, A. (2018). Tree-based global model tests for polytomous Rasch models. *Educational and Psychological Measurement*, *78*(1), 128–166.

33. Lin, X. E., & Boutros, P. (2019). NNLM: a package for fast and versatile nonnegative matrix factorization. Retrieved from https://github.com/linxihui/NNLM

34. Gaujoux, R., & Seoighe, C. (2010). A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*, *11*(1), 367.

35. Aggarwal, C. C. (2016). *Recommender systems: the textbook*. Switzerland: Springer International Publishing.

36. Aggarwal, C., & Parthasarathy, S. (2001). Mining massively incomplete data sets by conceptual reconstruction. In *Seventh ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 227–232).

37. Roux, J. le, Kameoka, H., Ono, N., Cheveign, A. de, & Sup, E. N. (2008). Computational auditory induction by missing data non-negative matrix factorization. In *ISCA Workshop on Statistical and Perceptual Audition (SAPA)* (pp. 1–6).

38. Taslaman, L., & Nilsson, B. (2012). A framework for regularized non-negative matrix factorization, with application to the analysis of gene expression data. *PLoS ONE*, *7*(11), 1–7.