Title

Statistical harmonization of versions of measures across studies using external data: Self-rated health and self-rated memory.

Authors

Wu, Yingyan
Hayes-Larson, Eleanor
Zhou, Yixuan
et al.

Publication Date

2025-02-01

DOI

10.1016/j.annepidem.2025.01.002

Peer reviewed

HHS Public Access

Author manuscript

Ann Epidemiol. Author manuscript; available in PMC 2025 March 05.

Published in final edited form as:

Ann Epidemiol. 2025 February ; 102: 86–90. doi:10.1016/j.annepidem.2025.01.002.

Statistical harmonization of versions of measures across studies using external data: self-rated health and self-rated memory

Yingyan Wu[1], **Eleanor Hayes-Larson**[1], **Yixuan Zhou**[1,2], **Vincent Bouteloup**[3,4], **Scott C. Zimmerman**[5], **Anna M. Pederson**[5,6], **Vincent Planche**[7,8], **Marissa J. Seamans**[1], **Daniel Westreich**[9], **M. Maria Glymour**[5,6], **Laura E. Gibbons**[10], **Carole Dufouil**[3,4], **Elizabeth Rose Mayeda**[1]

[1].Department of Epidemiology, University of California, Los Angeles Fielding School of Public Health, CA, USA

[2].Department of Biostatistics, University of California, Los Angeles Fielding School of Public Health, CA, USA

[3].Univ. Bordeaux, Inserm, Bordeaux Population Health, UMR1219, Bordeaux, France

[4].CIC 1401 EC, Pôle Santé Publique; CHU de Bordeaux, France

[5].Department of Epidemiology and Biostatistics, University of California, San Francisco, CA, USA

[6].Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA

[7].Univ. Bordeaux, CNRS, Institut des Maladies Neurodégénératives, UMR 5293, Bordeaux, France

[8].Pôle de Neurosciences Cliniques, Centre Mémoire de Ressources et de Recherche, CHU de Bordeaux, France

[9].Department of Epidemiology, University of North Carolina Gillings School of Global Public Health, Chapel Hill, North Carolina, USA

[10].Department of Medicine, School of Medicine, University of Washington, Seattle, WA, USA

Abstract

Purpose: Harmonizing variables for constructs measured differently across studies is essential for comparing, combining, and generalizing results. We developed and fielded a brief survey to harmonize Likert and continuous versions of measures for two constructs, self-rated health and self-rated memory, for use in studies of French older adults.

Corresponding author: Elizabeth Rose Mayeda PhD, MPH, 650 Charles E Young Dr S, CHS 46-070B, Los Angeles, CA 90095, mayeda@g.ucla.edu.

Ethics approval: This study was approved by the Data Protection Officer of the Bordeaux University Hospital, an Institutional Review Board evaluation was not required (MR004). The University of California, Los Angeles IRB classified this study as not human subjects research and determined that IRB review was not required.

Conflict of interest: None declared.

Methods: We recruited 300 participants from a French memory clinic in 2023 to answer both the Likert and continuous versions of self-rated health and self-rated memory questions. For each construct, we predicted responses to the Likert version with multinomial and ordinal logistic models, varying specifications of continuous version responses (linear or spline) and covariate sets (question order, age, sex/gender, and interactions between the continuous version and covariates). We also implemented a percentiles-based crosswalk sensitivity analysis. We compared Cohen's weighted kappa values to identify the best statistical harmonization approach.

Results: In the final models [multinomial models with continuous version spline, question order (self-rated memory model only), age, sex/gender, and interactions between the continuous version and covariates], weighted kappa values were 0.61 for self-rated health and 0.60 for self-rated memory, reflecting moderate agreement.

Conclusions: Primary data collection feasibly facilitates statistical harmonization of variables for constructs measured differently across studies.

Keywords

Statistical harmonization; Measurement; Primary data collection

Introduction

Important constructs are often measured differently across studies. Variability in versions of measures across studies makes it difficult to compare, combine (e.g., meta-analyze), and generalize study results, because variation in results could reflect either true differences or measurement heterogeneity. In some cases, measures can be harmonized across studies by coarsening categorical variables or applying thresholds for continuous variables based on clinical or theoretical justification (e.g., applying body mass index categories captured in one study to continuous body mass index captured in another study).[1] However, in cases without straightforward direct harmonization (e.g., when there is no clear threshold for categorizing a continuous variable), statistical harmonization can facilitate developing a "crosswalk" that aligns corresponding values from one version of the measure onto the other.[2,3] In contrast to direct harmonization, statistical harmonization is usually simultaneously theory-informed and data-driven.[3] Statistical harmonization generally requires some overlapping data (i.e., data collected from the same people on both versions of measures).

In this report, we describe one approach, including data collection and statistical methodology, for using statistical harmonization to develop a crosswalk between two versions of measures of self-rated health and self-rated memory. We were motivated by the need for harmonized measures in applied work (forthcoming) extending findings from a French memory clinic cohort to the French older adult population.[4] Self-rated health and self-rated memory were each measured using continuous scales in the clinic-based study and Likert scales in the nationally representative study, and there were no clear thresholds to directly harmonize the different versions of the measures, necessitating statistical harmonization to develop a crosswalk. The goal of the present study was to develop a crosswalk for these measures using a newly recruited external sample and demonstrate feasibility of primary data collection to facilitate crosswalk development. Although our goal

was to use the harmonized variables for generalizability analyses, we anticipate that the approach we present could be used to harmonize other variables for various applications.

Methods

Overview of approach

Briefly, our approach involved (1) identification of a study sample for collecting external data and study approval; (2) data collection; (3) data analysis to develop the crosswalk, including both modeling and percentile-based approaches; (4) evaluation of the crosswalk using Cohen's weighted kappa.

Study population and design

To obtain overlapping data (responses to both versions of measures of each construct from the same people), we developed and fielded a brief survey with sample size based on power calculations for multinomial logistic models.[5] Participants were patients and family caregivers visiting the Centre Mémoire Ressources Recherche of Bordeaux Hospital, a memory clinic in Bordeaux, France, from March to September 2023. The clinic was chosen because we expected the crosswalk to be similar for participants recruited at this clinic as the population in which the crosswalk would be applied in separate work (French older adults, including those consulting for mild cognitive problems). Eligibility criteria included age 60+ years and willingness and availability to complete the survey at the clinic. We excluded individuals who did not consent, could not understand the survey questions, were under guardianship, tutorship, or deprived of liberty by a juridical or administrative decision, or could not count or read.

Participants were asked to complete a tablet-based anonymous survey in French while waiting for a clinic visit via REDCap, an online survey web application.[6] The six survey questions included participant age group (60–69, 70–79, 80–89, or ≥90 years), sex/gender, and both the continuous and Likert versions of self-rated health and self-rated memory. The order in which the versions were presented ("question order") was randomized.

Staff provided potential participants with study objectives and their right to refuse participation, ensured inclusion criteria were met, and obtained oral consent from participants. This study was classified as not human subjects research by the University of California, Los Angeles Institutional Review Board and was approved by Bordeaux University Hospital.

Measures for harmonization

Self-rated health—The continuous version of self-rated health was a single item from the EQ-5D, a measure of health-related quality of life developed by the EuroQol Group.[7] Participants were asked to rate their health from 0 to 100 using a visual analog scale shown on the tablet screen, with 0 as the worst health and 100 as the best health the participant could imagine. The Likert version was a single item that asked participants to rate their health as "poor," "fair," "good," "very good," or "excellent." English translations of the prompts are shown in Figure S1.

Self-rated memory—The continuous version of self-rated memory was a single item that asked participants to rate their level of memory concern from 0 (not having any concern) to 10 (having concerns at the maximum level) using a visual analogic scale shown on the tablet screen. The Likert version was a single item that asked participants to rate their memory as "poor," "fair," "good," "very good," or "excellent." English translations of the prompts are shown in Figure S1.

Statistical analysis

We used the statistical approach described below to harmonize the two versions of each measure. Broadly, we fit several models using responses to the continuous version to predict responses to the Likert version and compared model performance and fit. We used the Likert version as the dependent variable because the Likert version has less information, so we expected our capacity to recover the continuous version would be more limited. Based on the small number of responses (≤30) at extremes of the Likert scale ("excellent" and "poor"), in our main analysis, Likert versions were collapsed into 3 categories: "excellent/very good," "good," and "fair/poor."

Using this 3-category Likert dependent variable, we estimated multinomial and ordinal logistic regression models, each with 12 specifications of predictor variables. Model 1 included only a linear term for the continuous version of the measure. Model 2 used a restricted cubic spline for the continuous version with internal knots at the 10^{th}, 50^{th}, and 90^{th} percentiles, and boundary knots at the 5^{th} and 95^{th} percentiles. To test whether question order affected response patterns, in Models 3–4, we added an indicator for question order and an interaction term between the continuous version of the measure and this indicator to Models 1–2. In Models 5–8, we added age group and sex/gender to Models 1–4. In Models 9–12, we added interaction terms between the continuous version of the measure and both age group and sex/gender to Models 5–8.

We generated a predicted value of the 3-category Likert response from each model for each construct. To evaluate model performance, we calculated Cohen's weighted kappa using quadratic weighting comparing the predicted and observed 3-category Likert responses.[8] For each construct, we chose the model with the highest kappa value as the final model to harmonize the continuous and Likert versions of measures.

We conducted several sensitivity analyses. First, we repeated analyses using 5-category Likert responses and calculated linear weighted kappa values to compare our results to a benchmark measure of agreement: test-retest reliability for self-rated health.[9] Second, a few participants (outlying points in Figure 1) may have been confused by the reverse coding of the continuous version of self-rated memory (higher scores indicated greater memory concerns/poorer memory) and Likert version of self-rated memory (higher scores indicated better memory). We dropped outliers, defined as observations where the continuous version values were more than 1.5 times the interquartile range below the first quartile or above the third quartile for each category of the Likert scale version.[10] Finally, we used a non-parametric percentile-based approach to calculate crosswalks, where we used question order randomization to calculate a crosswalk based on percentiles of scores for the first question version answered by each participant. Specifically, we calculated percentiles of continuous

version scores for participants who answered the continuous version first and percentiles of Likert version scores for participants who answered the Likert version first. We then used the percentile rankings to crosswalk between the versions and calculated weighted kappa values comparing the predicted and observed Likert responses. All analyses used R v4.1.3; code is on GitHub: https://github.com/Mayeda-Research-Group/bordeaux_survey_crosswalk.

Results

Of 314 survey responses, 300 were complete and comprised the analytic sample. Most (57%) participants were age 70–79; 50% identified as women (Table 1). Generally, higher continuous scores for self-rated health (indicating better health) and lower continuous scores for self-rated memory (indicating fewer memory concerns) tracked with Likert scores indicating better self-rated health and self-rated memory, respectively, although there was substantial overlap in continuous scores across Likert categories (Figure 1).

We compared weighted kappa statistics across the 12 models for each construct, excluding a few complex ordinal models that did not converge. Weighted kappa values and confidence intervals were similar across models and reflected moderate agreement[8] (self-rated health range: 0.56–0.61; self-rated memory range: 0.51–0.60; Table 2).

Our final model for self-rated health was the multinomial model with a spline for the continuous version, age group, sex/gender, and two-way interaction terms between the continuous version and the other predictors (multinomial model 10, weighted kappa=0.61 [95% CI: 0.53–0.69]). Our final model for self-rated memory was the multinomial model with a spline for the continuous version, question order, age group, sex/gender, and two-way interaction terms between the continuous version and other predictors (multinomial model 12, weighted kappa=0.60 [95% CI: 0.52–0.68]). Table S1 displays coefficients of the final models. Distributions of residuals from final models were similar across values of the continuous versions (Figure S2).

Compared to the main analyses, weighted kappa values were slightly lower in sensitivity analyses using the 5-category Likert variable (Table S2) and slightly higher after dropping outliers (self-rated health n=5, self-rated memory n=2) (Table S3). Weighted kappa values from the percentile-based approach were similar to the model-based approach (Table S4).

Discussion

In this study, we demonstrated one approach to obtaining harmonized measures of a construct. We developed and fielded a brief survey to collect responses from the same people for two versions of measures (one Likert scale, one continuous) of self-rated health and self-rated memory and used statistical models and percentiles to obtain predicted Likert version responses from the continuous version responses. This builds on prior harmonization work by combining an external sample with multiple analytic approaches.

Our harmonization approach was straightforward to implement and largely successful. Responses to the Likert versions of self-rated health and self-rated memory were distributed across “excellent/very good,” “good,” and “fair/poor,” and responses to the two versions

of measures tended to correlate as expected. Weighted kappa statistics assessing agreement between predicted and observed 3-category Likert responses were in a range considered to be moderate.[8] However, agreement between two versions of measures is unlikely to be higher than test-retest reliability, and kappa values for self-rated health in our study were similar to data from a National Health and Nutrition Examination Survey study of US adults, which reported moderate test-retest reliability (linear weighted kappa = 0.56).[9] While test-retest reliability likely varies to some extent across populations, this provides a helpful benchmark for evaluating kappa values in our study.

A strength of our approach is that all respondents answered questions for two versions of self-rated health and self-rated memory, with randomized question order. We evaluated several models and a percentile-based approach to harmonize versions of measures; weighted kappa values and confidence intervals were similar. A limitation of our study was that all respondents were from a single memory clinic; as a result, responses might not fully represent the joint distribution of responses for all potential populations of interest. Additionally, our survey was self-administered on a tablet; mode of administration could affect responses and the crosswalk. We assumed no heterogeneity in the joint distribution of continuous and Likert responses except by age and sex/gender; if other sources of heterogeneity are hypothesized, these variables would need to be included in primary data collection, and a larger sample size would likely be required to model the heterogeneity. Repeating two questions assessing the same construct may have influenced responses. We tried to eliminate the influence by randomizing question order and adjusting for question order and its interaction with the continuous version in models predicting Likert version responses. Additionally, kappa values from the percentile-based approach, which only used scores for the first question version answered by each participant, were very similar to results from regression models.

In conclusion, we demonstrated feasibility of external primary data collection as a potential solution to harmonize measures across studies. This study specifically focused on harmonizing versions of self-rated health and self-rated memory among French older adults, but our approach can be applied to harmonize versions of measures of other health-related constructs in other populations, as fielding small measurement studies is feasible in clinics or online. Our approach is also easily extended (e.g., with large enough samples, cross-validation and other modeling strategies, such as machine learning, could be applied, and inclusion of test-retest reliability in the study design could provide additional benchmarks for evaluating crosswalks). As such, we anticipate that the harmonization approach we present is likely to be straightforward and useful for many applications in which harmonized measures are necessary.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

We thank the following individuals and institutions for their expertise and assistance with data collection: Lauralee Meunier, Valérie Boilet, Delphine Jean, and the team at the *Centre Mémoire Ressources Recherche* of Bordeaux University Hospital

This work was conducted using the MRC Dementias Platform UK (DPUK). DPUK is a Public Private Partnership funded by the Medical Research Council (MR/L023784/1 and MR/009076/1). For further information on this resource visit www.dementiasplatform.uk.

Funding:

This work was supported by National Institute on Aging grant numbers R56AG069126 and R01AG072681, Eunice Kennedy Shriver National Institute of Child Health and Human Development grant number P2C-HD041022, and the University of California, Los Angeles Hellman Fellows Fund.

References

1. Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ. 2006 May 6;332(7549):1080.1. [PubMed: 16675816]
2. Chan KS, Gross AL, Pezzin LE, Brandt J, Kasper JD. Harmonizing Measures of Cognitive Performance Across International Surveys of Aging Using Item Response Theory. J Aging Health. 2015 Dec;27(8):1392–1414. [PubMed: 26526748]
3. Kołczyńska M Combining multiple survey sources: A reproducible workflow and toolbox for survey data harmonization. Methodol Innov. SAGE Publications Ltd; 2022 Mar 1;15(1):62–72.
4. Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. Generalizing Study Results: A Potential Outcomes Perspective. Epidemiology. 2017 Jul;28(4):553–561. [PubMed: 28346267]
5. Jong VMT de, Eijkemans MJC, Calster B van, et al. Sample size considerations and predictive performance of multinomial logistic prediction models. Stat Med. 2019 Apr 30;38(9):1601–1619. [PubMed: 30614028]
6. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009 Apr 1;42(2):377–381. [PubMed: 18929686]
7. Rabin R, Charro F de. EQ-5D: a measure of health status from the EuroQol Group. Ann Med. 2001 Jul;33(5):337–343. [PubMed: 11491192]
8. McHugh ML. Interrater reliability: the kappa statistic. Biochem Medica. 2012 Oct 15;22(3):276–282.
9. Zajacova A, Dowd JB. Reliability of Self-rated Health in US Adults. Am J Epidemiol. 2011 Oct 15;174(8):977–983. [PubMed: 21890836]
10. Dash ChSK, Behera AK, Dehuri S, Ghosh A. An outliers detection and elimination framework in classification task of data mining. Decis Anal J. 2023 Mar 1;6:100164.

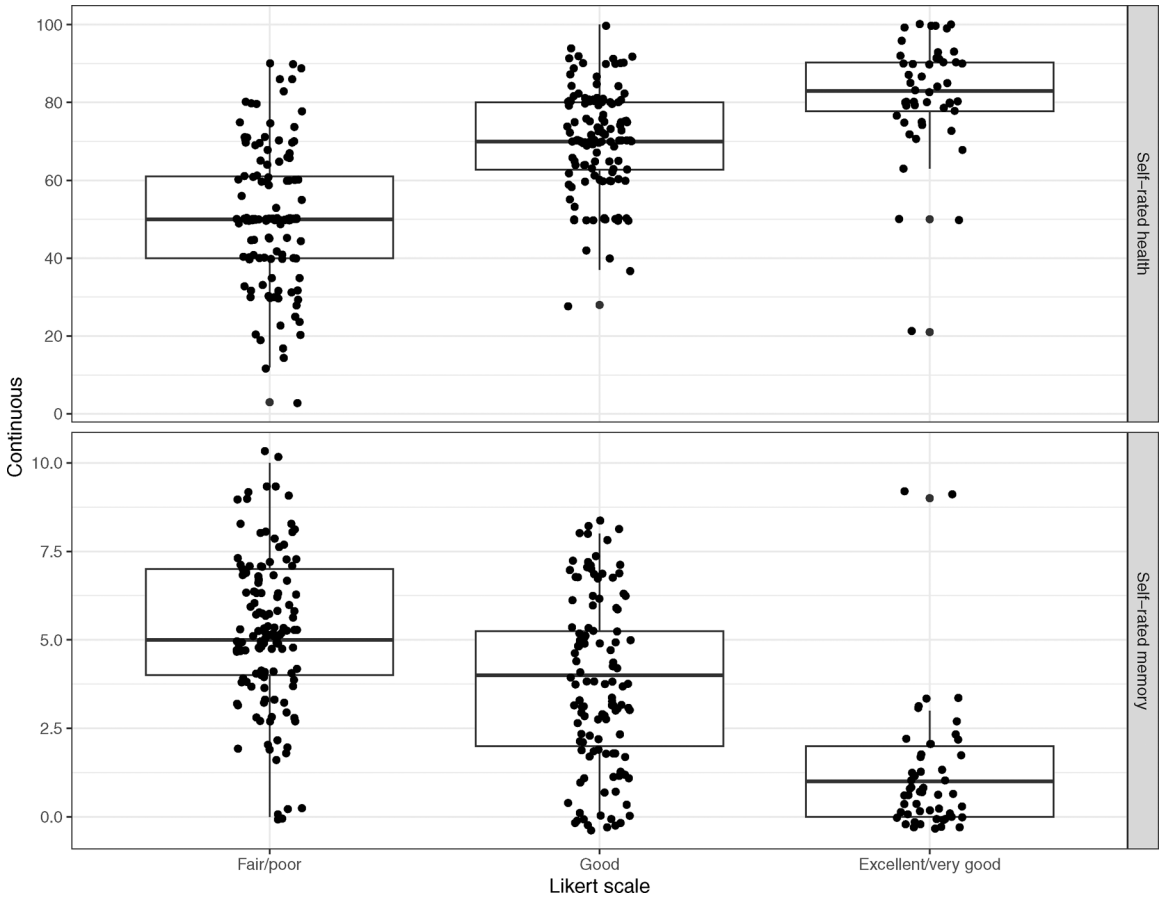

Figure 1.

Joint distributions of responses to versions of measures of self-rated health and self-rated memory shown by box scatter plot with jittering. Likert response categories are organized along the horizontal axis and continuous response values are shown along the vertical axis. For self-rated health, the range is 0–100; higher scores represent better health. For self-rated memory, the range is 0–10; higher scores represent greater memory concerns.

Table 1.

Characteristics of survey participants.

	***N* = 300**
Age group	
60 – 69	105 (35%)
70 – 79	170 (57%)
80 – 89	25 (8%)
Women	151 (50%)
Question order	
Continuous version first	149 (50%)
Likert version first	151 (50%)
Self-rated health (continuous, range 0–100, higher scores represent better health)	
Mean (SD)	63.9 (19.3)
Median (Q1, Q3)	67 (50, 80)
Self-rated health (Likert)	
Excellent	11 (4%)
Very good	37 (12%)
Good	124 (41%)
Fair	98 (33%)
Poor	30 (10%)
Self-rated memory (continuous, range 0–10, higher scores represent greater concerns)	
Mean (SD)	3.9 (2.6)
Median (Q1, Q3)	4 (2, 6)
Self-rated memory (Likert)	
Excellent	8 (3%)
Very good	45 (15%)
Good	116 (39%)
Fair	109 (36%)
Poor	22 (7%)

Abbreviations: SD, standard deviation; Q1, first quartile; Q3, third quartile

Table 2.

Cohen's weighted kappa statistics (95% confidence intervals) comparing predicted and observed 3-category measures of self-rated health and self-rated memory.

	Self-rated health weighted kappa	Self-rated memory weighted kappa
Multinomial model for 3-category outcome		
M1	0.58 (0.51, 0.66)	0.59 (0.51, 0.67)
M2	0.57 (0.49, 0.66)	0.59 (0.51, 0.67)
M3	0.56 (0.48, 0.64)	0.59 (0.51, 0.67)
M4	0.58 (0.49, 0.66)	0.59 (0.50, 0.67)
M5	0.57 (0.49, 0.65)	0.52 (0.43, 0.61)
M6	0.57 (0.48, 0.66)	0.54 (0.45, 0.63)
M7	0.59 (0.51, 0.67)	0.53 (0.44, 0.62)
M8	0.58 (0.49, 0.67)	0.58 (0.49, 0.66)
M9	0.56 (0.48, 0.65)	0.52 (0.43, 0.61)
M10	0.61 (0.53, 0.69)	0.59 (0.51, 0.67)
M11	0.57 (0.48, 0.65)	0.53 (0.43, 0.62)
M12	0.60 (0.51, 0.68)	0.60 (0.52, 0.68)
Ordinal model for 3-category outcome		
M1	0.58 (0.50, 0.66)	0.51 (0.43, 0.60)
M2	0.58 (0.50, 0.66)	0.54 (0.46, 0.62)
M3	0.56 (0.48, 0.64)	0.51 (0.43, 0.60)
M4	0.57 (0.49, 0.66)	DNC
M5	0.57 (0.49, 0.66)	0.55 (0.48, 0.63)
M6	0.58 (0.50, 0.67)	0.56 (0.48, 0.64)
M7	0.57 (0.49, 0.65)	0.56 (0.48, 0.63)
M8	0.58 (0.49, 0.66)	DNC
M9	DNC	0.56 (0.49, 0.64)
M10	DNC	DNC
M11	DNC	0.55 (0.48, 0.63)
M12	DNC	DNC

M1: linear term for continuous version

M2: restricted cubic spline for continuous version with internal knots at the 10^{th}, 50th, and 90^{th} percentiles and boundary knots at the 5^{th} and 95^{th} percentiles

M3: linear term for the continuous version, question order, and an interaction term between question order and the linear term for the continuous measure

M4: restricted cubic spline for continuous version with internal knots at the 10^{th}, 50th, and 90^{th} percentiles and boundary knots at the 5^{th} and 95^{th} percentiles, question order, and an interaction term between question order and the spline for the continuous version

M5: M1 + age group + sex/gender

M6: M2 + age group + sex/gender

M7: M3 + age group + sex/gender

M8: M4 + age group + sex/gender

M9: M5 + age group*continuous version + sex/gender*continuous version

M10: M6 + age group*continuous version + sex/gender*continuous version

M11: M7 + age group*continuous version + sex/gender*continuous version

M12: M8 + age group*continuous version + sex/gender*continuous version

DNC: did not converge