

Supplementary File 2

Short overview of the statistical packages used in the analysis

Mplus8.4

Mplus8.4 is a comprehensive statistical modeling program that provides a vast array of models, estimators, and algorithms to researchers. With its user-friendly interface and graphical displays, Mplus facilitates the analysis of cross-sectional and longitudinal data, as well as single-level and multilevel data. It is adept at handling data that come from different populations with either observed or unobserved heterogeneity, and managing datasets with missing values. Mplus is a versatile software that integrates the SEM framework for both single and multiple-group analyses. It supports models incorporating continuous latent variables, categorical latent variables, or a combination of both, and observed outcome variables may span a range of types, including continuous, binary, ordinal, nominal, censored, and counts, as well as mixtures of these types. The software provides a comprehensive range of estimators tailored to the model's design, including ML, MLR, MLM, WLS, WLSM, WLSMV and ULS. To assess model fit, Mplus offers several tests and indices. These include the chi-square test of model fit, the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), information criteria, the Root Mean Square Error of Approximation (RMSEA), and the Standardized Root Mean Square Residual (SRMR). Mplus is equipped for complex survey data analysis, providing diverse options for model estimation. It facilitates two distinct approaches to the analysis of such data. The first approach involves calculating standard errors and a chi-square test of model fit, with adjustments made for stratification, the non-independence of observations due to cluster sampling, and unequal probability of selection. The alternative method involves specifying a distinct model for each level of multilevel data, thus directly addressing the non-independence of observations resulting from cluster sampling (referred to as multilevel modeling). Both methods accommodate the use of sampling weights to adjust for unequal selection probabilities, where parameters are estimated through the maximization of a weighted loglikelihood function and standard errors are computed using a sandwich estimator. In the context of this study, the first approach is adopted. This is implemented by selecting the TYPE=COMPLEX option within the ANALYSIS command, which is further specified by incorporating STRATIFICATION, CLUSTER, and WEIGHT options in the VARIABLE command (Muthén and Muthén, 2017).

Stata18

Stata18 is an intricate and comprehensive statistical package tailored for data analysis, data management, and visualization. Beyond its extensive range of statistical functions, the package is equipped with a SEM module, featuring two primary commands: *sem* and *gsem*. The *sem* command is suited for fitting standard linear SEMs, where both responses and latent variables are continuous, and

the models employ linear regression. It assumes that variables are jointly distributed normally. On the other hand, the *gsem* command is designed for generalized SEMs, accommodating latent variables that are either continuous or categorical (not simultaneously), with responses that can be continuous, ordinal, binary, count, or multinomial. This command supports a variety of models, including linear regression, logit, or probit, without the requirement for a full joint normality assumption. While there is some overlap in their capabilities, the *sem* command offers unique features not available with *gsem*. These include the use of ADF estimator, the ability to handle observations with missing values (MLMV), goodness-of-fit statistics, modification indices, standardized estimates, and the capacity to model using summary-statistic data. Both commands facilitate multigroup analysis, hypothesis tests, and accommodate models with complex survey data. Interestingly, when *sem* and *gsem* can fit the same model, they produce the same or almost equal numerical solutions for parameters and standard errors. This similarity arises because both *sem* and *gsem* estimate the same mathematical model, albeit through different numerical mechanisms. Considering the advantages and limitations of the available options, we have decided to employ the *sem* command for our analysis (StataCorp, 2023).

Stata, and *sem* command, offers four types of parameter estimation for the naïve model: ML, QML, MLMV, and ADF. It provides several fit metrics: RMSEA, SRMR, CD (Coefficient of Determination), CFI, TLI, Chi-Squared Test, and Information Criteria (not applicable for ADF). However, when sample weights, and design variables are applied, the ADF method cannot be utilized. For ML, the SRMR and CD goodness of fit are displayed, and CD for MLMV. In practice, the SEM procedure for complex survey data can only be implemented using ML and MLMV estimation methods. StataCorp notes that the weight term specifies the significance attached to each observation, highlighting that not every command supports every type of weight. In the realm of Structural Equation Modeling, Stata18 offers three options for weighting variables: frequency weights, sampling weights [command: *pweights*], and importance weights, with detailed definitions provided in Stata's user guide. This article primarily focuses on the sampling weight option, which reflects the inverse probability of an observation's inclusion due to the sampling design. Moreover, it allows for the integration of sample design variables through the *svy* command. The foundational command structure for initiating complex survey data analysis in Stata18 is as follows: *svyset* cluster variable [*pweight*= weight variable], *strata*(stratification variable) (StataCorp, 2023).