# 

# **MUSE Coding Syntax Users Guide**

# Last Updated: 07/24/2020

# Overview

This users guide provides guidance for the syntax documents created for quantitative analysis in Phase 2 of the MUSE project. This includes SAS files, Mplus files, and R files. The SAS files primarily involve initial data cleaning, data set creation, and preliminary descriptive analysis. Mplus files include exploratory and confirmatory factor analysis for each scale. R files include graphical factor structure representations and item response theory graded response models for each of 15 scales. All syntax files and associated data files are located in the [**Phase 2 Quantitative Analysis folder on Box**](https://emory.app.box.com/folder/107817404317)**.**

# Table of Contents

[SAS Syntax Files 1](#_Toc42693127)

[Data Format Dictionary 1](#_Toc42693128)

[Data Cleaning 1](#_Toc42693129)

[EFA/CFA Split 2](#_Toc42693130)

[Descriptive Analysis 2](#_Toc42693131)

[Mplus Syntax 3](#_Toc42693132)

[Exploratory Factor Analysis 3](#_Toc42693133)

[Confirmatory Factor Analysis 3](#_Toc42693134)

[R Syntax 5](#_Toc42693135)

[Factor Structures 5](#_Toc42693136)

[Item Response Theory 5](#_Toc42693137)

# 

# SAS Syntax Files

In this phase of the MUSE analysis, SAS was used to (1) create a dictionary of defined formats for MUSE variables; (2) clean the initial collected data, including deleting observations flagged by the research team, overwriting administrative variables where identified by the research team, formatting and labeling all variables, reverse-scoring, collapsing, and combining relevant variables, and scoring preliminary versions of each scale; (3) create a random split of the collected data into a data set for exploratory factor analysis and a data set for confirmatory factor analysis; and (4) carry out an initial descriptive analysis of the cleaned and formatted data to support a preliminary report to the Bill and Melinda Gates Foundation.

## Data Format Dictionary

**File Path:** [Women's Empowerment in WASH 🡪 Data analysis 🡪 Phase 2 Quantitative Analysis 🡪 Syntax](https://emory.app.box.com/folder/101882632165)

**File Name(s):** Format - MUSE

**Overview:** This file contains a proc format statement that creates a set of 48 variable formats to be applied to MUSE data variables. This file will be referenced and run through the data cleaning file mention below and does not need to be run on its own. Each variable format includes a comment indicating which variables to which the format will be applied.

## Data Cleaning

**File Path:** [Women's Empowerment in WASH 🡪 Data analysis 🡪 Phase 2 Quantitative Analysis 🡪 Syntax](https://emory.app.box.com/folder/101882632165)

**File Name(s):** MUSE\_Cleaning\_2020-03-30

**Overview:** This data cleaning and formatting files contains several different sections. The first section involves administrative cleaning of the data, including (1) setting the pathway to the data and referencing the previously mentioned format file; (2) loading the initial, uncleaned and unformatted data set; (3) deleting observations a priori as determined by the research team; and (4) changing neighborhood and enumerator IDs where incorrect as determined a priori by the research team. After this section, a copy of the cleaned, but still unformatted, data set is exported for internal quality control and record keeping. The second section involves formatting and labeling the data, including (1) applying the loaded formats to applicable variables; and (2) labeling variables that will be used for preliminary analysis. Again, after this section, a copy of the data is exported for internal quality control and record keeping. The third section involves (1) calculating new variables; (2) combining/collapsing existing variables; and (3) reverse-scoring variables identified a priori by the research team. After this section, a copy of the data set is exported for internal quality control and record keeping. The fourth section involves scoring all preliminary scales and indices – this is done prior to any scale reduction process with all scale and index items. Scales are scored as averages, while indices are summed. Prior to scoring, all missing data (true missing, not applicable, I don’t know, etc.) are set to be SAS missing responses. The fifth section involves (1) splitting the data into baseline and re-test data sets; (2) calculating quintiles for the India and Uganda wealth index for the baseline data only; and (3) exporting and saving both final, cleaned and formatted baseline and re-test data sets as SAS data files and .csv data files. At each stage of analysis, wherever a new temporary data set is created, comments indicate how many variables and observations should be in the new data set. Comments are also used to indicate any places where a user may need to make modifications (e.g. changing a library pathway to one from their own computer).

## EFA/CFA Split

**File Path:** [Women's Empowerment in WASH 🡪 Data analysis 🡪 Phase 2 Quantitative Analysis 🡪 Syntax](https://emory.app.box.com/folder/101882632165)

**File Name(s):** MUSE\_EFA\_CFA\_Split\_2020-03-19

**Overview:** As part of scale development best practices, exploratory and confirmatory factor analyses should be carried out on separate data sets. For logistic purposes, instead of doing two separate rounds of data collection, data was collected at one time point and is randomly split into two equally sized data sets. This syntax file can be used on the initial raw, unformatted data set. It replicates the first data cleaning section mentioned above in “Data Cleaning”, reverse-scores applicable scale items, filters out re-test data to include only initial baseline responses, and then splits the file in two using proc surveyselect with samprate set to 0.50. The two data files are then exported as CSV files for further analysis in R and Mplus. At each stage of work done in this file, wherever a new temporary data set is created, comments indicate how many variables and observations should be in the new data set. Comments are also used to indicate any places where a user may need to make modifications (e.g. changing a library pathway to one from their own computer).

## Descriptive Analysis

**File Path:** [Women's Empowerment in WASH 🡪 Data analysis 🡪 Phase 2 Quantitative Analysis 🡪 Syntax](https://emory.app.box.com/folder/101882632165)

**File Name(s):** MUSE\_Descriptives\_2020-03-30

**Overview:** This data file was used to run descriptive statistics for a preliminary report on Phase 2 data for the Bill and Melinda Gates Foundation and partners in India and Uganda. The cleaned and formatted data output from the data cleaning syntax file is imported, and is then subset by country. Figures and tables are then generated as was determined a priori by the research team and saved to an external folder. Of note, this file utilizes SAS macros, to create and save tables, which are saved on Box in a sub-folder within the syntax folder linked above – *these macros were previously created, are open-source, and are not specifically a part of this MUSE project*. Comments in the syntax are used to indicate any places where a user may need to make modifications (e.g. changing a library pathway to one from their own computer to correctly load the data and/or macros).

# Mplus Syntax

For each of the 15 scales in Phase 2, Mplus was used to carry out exploratory and confirmatory factor analysis. Exploratory factor analysis (EFA) was done iteratively, with the research team reviewing and dropping items for a given scale each round. Additionally, EFA syntax includes a line to output scree plots, which were then analyzed within R (see “Factor Structure Visualization” section below), as the version of Mplus being used did include graphic capabilities within the application itself. After each preliminary scale was modified and shortened through EFA and item response theory, confirmatory factor analysis (CFA) was then carried out. For organization, EFA and CFA syntax is organized by domain and then by scale.

## Exploratory Factor Analysis

**File Path:** [Women's Empowerment in WASH 🡪 Data analysis 🡪 Phase 2 Quantitative Analysis 🡪 EFA](https://emory.app.box.com/folder/107611971284)

**File Name(s):** Within the above link, files are sorted into folders by domain (Agency, Institutional Structures, and Resources) and then by scale name. Each scale folder contains a syntax folder; the .inp files are Mplus syntax. As EFA was done iteratively as scale reduction decisions were made by the research team, there may be multiple Mplus files for each scale. Each file is labeled with the first initial of the domain, the scale name, the round of EFA, and the date last modified: E.g., A\_Lead\_R3\_2020-04-10.inp.

**Overview:** All Mplus EFA files were created with Mplus 8, Mplus Editor Version 1.7 on a Mac operating system. As such, Mplus did not have graphic capabilities so each EFA file also includes a plot command to create a .gh5 file that can then be read by R (see ‘Factor Structures” section below) to create a scree plot for determination of the optimal number of factors for each scale. Each file has 6 sections: (1) Title; (2) Data; (3) Variable; (4) Analysis; (5) Plot; and (6) Output. The variable section has 4 subsections of syntax: ‘Names’ defines the names for all variables in the data file; ‘Usevariables’, the variables being used for this round of analysis (as scale reduction decisions are made in an iterative fashion variable names are dropped from the list here – NOT from the list of all variables in the data set); ‘Categorical’ a list of the variables being used that are categorical (for MUSE, as all variables are categorical, this will be the same set of variables as Usevariables); and ‘Missing’ which defines all values of “.” as missing data. The analysis section indicates the type of analysis as EFA, with requested range of number of factors for each scale (e.g. “Type is EFA 5 7” runs 3 EFA analyses with 5, 6 and 7 factors). All files have WLSMV (weighted least square mean and variance adjusted) as the estimator and quartimin as the rotation. All files request PLOT2 in the plot line to create the .gh5 scree plot file mentioned above, and modification indices (modindices) as output.

## Confirmatory Factor Analysis

**File Path:** [Women's Empowerment in WASH 🡪 Data analysis 🡪 Phase 2 Quantitative Analysis 🡪 CFA](https://emory.app.box.com/folder/113839314846)

**File Name(s):** Within the above link, files are sorted into folders by domain (Agency, Institutional Structures, and Resources) and then by scale name. Each scale folder contains a syntax folder with an .inp file labeled with the first initial of the domain, the scale name, and the date last modified: E.g., A\_Lead\_2020-05-31.inp.

**Overview:** All Mplus CFA files were created on Mplus 8, Mplus Editor Version 1.7 on a Mac operating system. Each file has 6 sections: (1) Title; (2) Data; (3) Variable; (4) Analysis; (5) Model; and (6) Output. The variable section has 4 subsections of syntax: ‘Names’ defines the names for all variables in the data file; ‘Usevariables’, the variables being used for this round of analysis (as scale reduction decisions are made in an iterative fashion variable names are dropped from the list here – NOT from the list of all variables in the data set); ‘Categorical’ a list of the variables being used that are categorical (for MUSE, as all variables are categorical, this will be the same set of variables as Usevariables); and ‘Missing’ which defines all values of “.” as missing data. The analysis section indicates the type of analysis as ‘general’. All files have WLSMV (weighted least square mean and variance adjusted) as the estimator. The model section specifies the factor structure for the given scale as determined by the research team through the EFA/item reduction process. Each of n factors is labeled as fn and then BY statements are used to assign individual items to a given factor. E.g.:

f1 BY K2 K3 K7;

f2 BY K9 K10 K12;

f3 BY K14 K15 K16 K19;

All files request modification indices (modindices), standardized loadings (Standardized), and sample statistics (sampstat) in the output section.

# 

# R Syntax

For Phase 2 of MUSE analysis, R was used to (1) create determine the optimal number of factors and create visual representations of the factor structure of each scale (scree plots and parallel analysis); and (2) carry out item response theory (IRT) analyses for polytomous items using graded response models for each scale.

## Factor Structures

File Path: [Women's Empowerment in WASH 🡪 Data analysis 🡪 Phase 2 Quantitative Analysis 🡪 EFA](https://emory.app.box.com/folder/107611971284)

File Name(s): Within the above link, files are sorted into folders by domain (Agency, Institutional Structures, and Resources) and then by scale name. Each scale folder contains a syntax folder; the .R files are R syntax. Each file is labeled with the first initial of the domain, the scale name, and the date last modified: E.g., A\_Lead\_2020-04-07.R.

Overview: Each file contains 5 sections. The first loads a list of packages needed for analysis: [psych](https://cran.r-project.org/web/packages/psych/psych.pdf), [GPArotation](https://cran.r-project.org/web/packages/GPArotation/GPArotation.pdf), [nFactors](https://cran.r-project.org/web/packages/nFactors/nFactors.pdf), and [sjPlot](https://cran.r-project.org/web/packages/sjPlot/sjPlot.pdf). The commands to install these packages are currently inset as comments, along with instructions to set them as active commands when the syntax file is run on a new computer for the first time. The file then includes steps to load a [special package](https://www.bioconductor.org/) that will allow R to read Mplus graphic files, these should be run each time. A comment indicates where the file pathway will need to be changed if being run on a new computer operating system. The second section loads the data set for a given scale, subsets to only scale items if there are also index items, and then creates further subsets for each round of item reduction, if applicable. The third section creates a table of item-level indices (mean score, standard deviation, % missing, and skewness) for all items. The fourth section is the determination of the optimal number of factors, including (1) a scree plot based on the .gh5 files generated from Mplus, (2) a scree plot using R functions, and (3) parallel analysis using R functions. The fifth section generates citations for all packages used.

## Item Response Theory

File Path: [Women's Empowerment in WASH 🡪 Data analysis 🡪 Phase 2 Quantitative Analysis 🡪 IRT 🡪 Syntax](https://emory.app.box.com/folder/113838510823)

File Name(s):

Overview: Each file contains 4 sections. The first loads the package needed for analysis, [mirt](https://cran.r-project.org/web/packages/mirt/mirt.pdf). The command to install this package is currently inset as a comment, along with instructions to set it as an active command when the syntax file is run on a new computer for the first time. The second section loads the data set for a given scale, subsets to only scale items if there are also index items, and then creates further subsets for each round of item reduction, if applicable. The third section runs the IRT graded response model analysis. This includes (1) printing the discrimination and difficulty parameters for each item, (2) generating operating characteristic curves for each item, (3) generating individual information curves for each item, and (4) generating a total information curve for the whole scale. The fourth section generates a citation for the package used.