A Practical Guide for Analyzing Large-Scale Assessment Data Using Mplus: A Case Demonstration Using the Program for International Assessment of Adult Competencies Data

Takashi Yamashita ⓘ
University of Maryland, Baltimore County

Thomas J. Smith
Northern Illinois University

Phyllis A. Cummins
Miami University

In order to promote the use of increasingly available large-scale assessment data in education and expand the scope of analytic capabilities among applied researchers, this study provides step-by-step guidance, and practical examples of syntax and data analysis using Mplus. Concise overview and key unique aspects of large-scale assessment data from the 2012/2014 Program for International Assessment of Adult Competencies (PIAAC) are described. Using commonly-used statistical software including SAS and R, a simple macro program and syntax are developed to streamline the data preparation process. Then, two examples of structural equation models are demonstrated using Mplus. The suggested data preparation and analytic approaches can be immediately applicable to existing large-scale assessment data.

Keywords: *assessment data analysis; sampling weights; plausible values*

Introduction

This article provides practical guidance for analyzing large-scale assessment data using the structural equation and latent variable modeling software application—Mplus (Muthén & Muthén, 1998–2017). The term "large-scale assessment" generally refers to " . . . surveys of knowledge, skills, or behaviors in one or more given domains" (Kirsch & Lennon, 2017, p. 2). Starting in the early 1990s, data from several large-scale assessments have increasingly become available to researchers and practitioners in the field of education. In particular, the

Organization for Economic Cooperation and Development (OECD) has played a key role in conducting international data collection to systematically assess individuals across global communities from early childhood to adulthood, including individuals from various formal educational settings (elementary, secondary, and postsecondary education; McFarland et al., 2018). For example, the Programme for International Student Assessment (PISA), Progress in International Reading Literacy Study (PIRLS), Trends in International Mathematics and Science Study (TIMSS), and the Program for the International Assessment of Adult Competencies (PIAAC) are some current assessments that are periodically conducted (Cresswell et al., 2015). These data sets are all publicly available and useful for describing education outcomes as well as cognitive skill domains (e.g., literacy, numeracy) at each developmental stage, understanding trends over time, and identifying differences among diverse OECD countries. Additionally, in combination with the background and contextual data collected in each large-scale assessment, education researchers have opportunities to address complex research questions with nationally and internationally representative samples. Moreover, large-scale assessment data can serve as valuable resources to facilitate decision making among policymakers, school administrators, and educators.

Despite the fact that large-scale assessment data are indeed useful and carry great potential for both research and practice, analysis of such data requires specialized skills in areas such as database management, statistics, and statistical programming. Also, both the large sizes of these data sets and the methodological complexity often become barriers to applied researchers. Namely, researchers should comprehend the sampling methods, sampling weights, plausible values, and variance estimation and should learn how to incorporate these key features of large-scale assessment data into their analyses.

Unique Features of Large-Scale Assessment Data

As an applied example, we use the 2012/2014 U.S. PIAAC Survey of Adult Skills public use file (PUF) data. This example is generalizable, and similar analytic strategies could be used with other large-scale assessment data such as PISA, PIRLS, and TIMSS. PIAAC was designed as a successor to earlier basic skill assessments such as the International Adult Literacy Survey conducted from 1994 to 1998 (Murray et al., 1998) and the Survey of Adult Literacy and Lifeskills conducted from 2003 to 2008 (National Center for Education Statistics [NCES], n.d.). The PIAAC target population consists of community-dwelling adults aged 16 years and older. In 2012 (first round) and 2014 (second round), a total of 34 countries participated in the PIAAC.

The sampling strategies vary slightly, but data for all countries were collected using multistage sampling methods per the technical standard and guidelines provided by OECD. For example, the United States employed its census data to determine sampling units (e.g., county) and develop the sampling

frames, which were intended to cover 95% or more of the target population (OECD, 2016). In the second round, the United States oversampled younger persons (age 16–34 years), unemployed, and older (age 66–74 years) adults. The 2012/2014 U.S. PIAAC data have a total of approximately 8,700 cases. The sampling weights were created by the developers based on this complex sampling design and are provided in the PIAAC data file (as variable SPFWT0). Data analysis incorporating the sampling weights adjusts for non-response bias and oversampling and estimates nationally representative parameters (e.g., average literacy skills).

For the data from both the computer- and paper-based instruments, the PIAAC adopted sophisticated skill assessment and estimation strategies using multistage adaptive testing as well as item response theory (IRT). In the multistage adaptive testing environment, respondents are not required to complete all survey items but only a subset of sequentially and systematically assigned assessment items. Based on the respondent's characteristics and performance on three skill assessments—literacy, numeracy, and problem-solving in technology-rich environments—10 sets of plausible values were estimated for each of these skill proficiencies from a specific statistical model, and these plausible values are provided in the PIAAC data. These plausible values or skill scores range from 0 to 500 points. In any analysis that uses at least one of these three skill proficiency measures, the use of all 10 sets of plausible values is recommended for the variance estimation of skill proficiency. Additional technical details pertaining to plausible values have been published elsewhere (Kirsch & Lennon, 2017; OECD, 2016; Rutkowski et al., 2010).

Another key consideration when using large-scale assessment data involves variance estimation. To estimate the variance and/or standard errors of variables of interest in PIAAC, the use of the supplied replicate weights is one recommended approach, although other approaches are available. These replicate weights were created based on the specific sampling strategies (i.e., multistage sampling) used in each country, and statistical analysis should apply appropriate estimation methods (OECD, 2016). In the case of the PIAAC U.S. public use data, 80 sets of replicate weights are provided and estimation requires the paired jackknife (also referred to as "jackknife 2") method, which pairs two subsamples from each primary sampling unit (i.e., geographic region such as county or census block) to form variance strata. Essentially, the jackknife 2 method systematically selects two samples from a larger cluster of cases, estimates the statistic of interest (e.g., mean literacy skill score) then repeats the process and empirically constructs the sampling distribution of the desired statistic along with its estimated standard error. The estimated standard errors, then, consist of two components: the variation estimated by the jackknife procedure and the variation across the 10 sets of plausible values. More detailed description of the replicate weights and relevant estimations in the large-scale assessment data has been published elsewhere (OECD, 2009; Rutkowski et al., 2010).

About the Tool

Fortunately, several analytic tools are available publicly to researchers and practitioners to analyze large-scale assessment data appropriately. The use of sampling weights, plausible values, and variance estimation procedures—the particulars of which are unique to each data set—is automated in most of these tools. Here, we briefly introduce two commonly used tools that accomplish these tasks. The first tool, the International Data Explorer (IDE) is a web-based application that produces representative statistics in table or figure format, based on the user inputs. The IDE can be accessed from the OECD (http://www.oecd.org/skills/piaac/publicdataandanalysis/) and the NCES (https://nces.ed.gov/surveys/international/ide/). Additionally, IDE has built-in statistical test functions for assessing bivariate significance and conducting regression analysis. For example, one may use the IDE tool to estimate the average literacy skill score of adults aged 16 to 65 years in the United States and fit a regression model with gender as a predictor of these literacy skill scores. IDE is particularly useful for international comparisons because the international database already has been constructed, and analytic procedures including the use of sampling weights, plausible values, and variance estimation are automated.

Another useful analytic tool for the large-scale assessment data such as PIAAC is the International Database (IDB) Analyzer (International Association for the Evaluation of Educational Achievement, 2016), which is a free application available from IEA (https://www.iea.nl/data-tools/tools) and OECD (http://www.oecd.org/skills/piaac/publicdataandanalysis/). The IDB Analyzer can generate macro programs for use in commercial statistical packages including SPSS (IBM Corp, 2016) and SAS (SAS Institute Inc., 2002–2012). The macro program merges large-scale assessment data from selected countries, estimates representative statistics, and conducts statistical analysis (specifically, bivariate tests and regression analysis), taking into account sampling weights, replicate weights, and plausible values. The IDB Analyzer is applicable not only to the PIAAC data but also to PISA, PIRLS, TIMSS, and many other large-scale assessment data in education. At the time of this study, the latest version of IDB Analyzer can fit a variety of generalized linear models such as proportional odds ordinal logistic regression and multinomial logistic regression models. For those who may need basic training in the use of IDE and/or the IDB Analyzer, online training resources are available from the OECD (http://www.oecd.org/skills/piaac/publicdataandanalysis/) and IEA (https://www.iea.nl/research-services/training#section-200). Additionally, it should be noted that other analytic tools to analyze large-scale data sets also are available, such as the *repest* package in Stata (Avvisati & Keslair, 2019) and the *intsvy* package in R (Caro & Biecek, 2017).

Tools such as the IDE and the IDB Analyzer certainly enable applied researchers to conduct analysis of large-scale assessment data using appropriate methods. At the same time, the scope of analysis is limited within these tools.

Although the IDE application and the IDB Analyzer are periodically updated with new analytic functions, provisions for other analytic methods of interest may not necessarily be available. In this respect, more flexible tools such as *repest* and *intsvy* packages can expand the scope of analysis in specific statistical packages. In addition, it is possible to use commercial software such as SPSS and SAS without the IDB Analyzer or other macro programs. Yet, advanced survey data analysis and programming skills are required to carry out estimation and statistical analysis using the sampling weights, replicate weights, and plausible values. Each statistical package has unique strengths and specific procedures (e.g., required data format) to appropriately analyze large-scale assessment data.

In this regard, to date, software-specific guidance for the applied education researcher who is interested in large-scale assessment data is limited. Therefore, the current study seeks to provide practical guidance and applied exemplar analyses of PIAAC data using the latent variable modeling application—Mplus (Muthén & Muthén, 1998–2017). Mplus is one of the most popular statistical packages for latent variable modeling, including structural equation modeling (SEM). SEM allows complex modeling strategies such as path analysis, mediation–moderation analysis, confirmatory factor analysis, and multivariate regression models with latent variables (Kline, 2016). We focus on the Mplus software application for several reasons. First, Mplus is capable of modeling non-normal or categorical outcomes both as observed and latent variables. Also, Mplus can fit a variety of increasingly popular SEMs including multigroup analysis, latent class analysis, latent profile analysis, and growth mixture models (Wang & Wang, 2012). Finally, Mplus continually adds new analytic functions (e.g., dynamic SEM, see Asparouhov et al., 2018). Many studies collecting large-scale assessment data are ongoing, and more data are continually becoming available. The capacity afforded by applications such as Mplus to carry out highly sophisticated analyses will help to advance research in education. For readers who are interested in more technical details and applications of specific SEM in Mplus using PIAAC, Scherer (2020) provides extensive illustrations in the international context and valuable methodological resources. This study instead provides a tutorial to analyze large-scale assessment data using the institutionally provided plausible values, sampling weights, and jackknife replicate weights as well as to verify the specialized functionality of Mplus to conduct these analyses in a way that is consistent with the recommended procedures from the institutions.

Applications

Empirical Examples

In this section, we demonstrate how to prepare the PIAAC data for analysis in Mplus. To use the sampling weights, replicate weights, and plausible values, we present five steps to prepare the data. We will use a simple example that

examines the association between literacy skills and motivation to learn in adult populations aged 25 years and older in the United States.

Data

Data were derived from the 2012/2014 PIAAC PUF. In the analyses presented here, the sample was limited to those aged 25 years and older in order to focus on the typical postsecondary education life stages. Considering only the set of variables used in the analyses, and after excluding those respondents with no valid data for these variables, the final sample size was 6,632. Available data from respondents with partially missing values were incorporated into the model estimation through the use of full information maximum likelihood (FIML) estimation (Arbuckle, 1996).

Models

Example #1—Structural equation model examining how motivation to learn is predicted by literacy skill.

Dependent variable: Motivation to learn. Per previous work by Gorges at al. (2016), four survey items from the PIAAC background questionnaire were used as indicators of a latent motivation to learn construct. The survey items consisted of four statements ("I like learning new things," "I like to get to the bottom of difficult things," "I like to figure out how different ideas fit together," and "If I don't understand something, I look for additional information to make it clear"), where each statement was associated with response options of 1 = *not at all* to 5 = *to a very high extent*. The measurement model for the latent readiness to learn outcome was assessed using CFA (Brown, 2014; Kline, 2016; Wang & Wang, 2012) as described in the statistical analysis section.

Independent variable: Literacy skills. Ten sets of plausible values for literacy skills consisting of scores ranging from 0 to 500. Higher scores indicate higher literacy proficiency.

Covariates: Age-group (in 5-year increments), gender (female vs. male), educational attainment (college or higher vs. less than college education), and self-rated health (1 = *poor* to 5 = *excellent*) were included in the analysis. Note that these variables were chosen as exemplar covariates and are not necessarily an exhaustive list of theoretically relevant covariates.

Example #2—Structural equation model examining how literacy skills are predicted by demographic characteristics.

Dependent variable: Literacy skills (see the description in Example #1).

Independent variables: Age-group, gender, race/ethnicity (Black, Hispanic, and Others [vs. White]), educational attainment, and self-rated health (see the description in Example #1).

Data Preparation

In this section, we present a suggested way to prepare a PIAAC data file in SAS or SPSS format for analysis in Mplus. Exemplar SAS and R syntax for implementing this is provided in the Appendices in the online version of the journal. The data preparation proceeds in five steps:

    1. Recode the variables of interest and check the resultant coding.

Variables in the PIAAC data set are recoded, including those measuring the motivation to learn construct, age, gender, race/ethnicity, educational attainment, and self-rated health. The coding results are checked by generating and examining descriptive statistics (e.g., minimum, maximum, mean) and graphical displays (e.g., histograms).

    2. Create data subset if needed

A subset of the PIAAC data containing the selected variables is created to minimize the necessary computation time and the potential for syntax error (e.g., errors that can occur due to listing a very large number of variable names in Mplus). Additionally, in this step, a subset of the observations in the data set is selected, consisting of adults aged 25 years and older. If analysis of a particular subset of observation is desired, this selection of observations should be done in the data setup stage because Mplus does not allow selection of subsamples when replicate weights are used. It should be noted, however, that caution should be taken when creating small and unique subsets of data, as these can result in incompatibilities with the original sampling weights (e.g., see Gelman, 2007).

    3. Recode any missing values

When using input data files that are in "free format," missing data values must be explicitly coded (i.e., user-specified) as numeric values for use in Mplus. In our illustration, missing data values—originally coded as N, D, R, or "system missing" (i.e., ".") in SAS or SPSS—have been recoded as -9999 for use in Mplus. Explicitly assigning subtypes of missing values to a specified code (-9999) is a good practice in this process. On a relevant note, it is possible to manually recode all missing values when, for example, modeling a small number of variables and/or data with relatively few cases. However, we recommend using an automated procedure such as provided in the SAS or R syntax to avoid inadvertent data coding errors.

    4. Create 10 data sets, where each consists of the variables of interest plus one unique set of plausible values.

To fit models that use the PIAAC plausible values in Mplus, 10 data sets—each containing the variables of interest in addition to one of the 10 sets of

plausible values—need to be created. For example, one data set generated by the supplied syntax includes all the model variables of interest (i.e., gender, education, self-rated health), the sampling and replicate weights, and the first literacy plausible value variable (PVLIT1). The second data set includes the same variables but substitutes the second literacy plausible value variable (PVLIT2). In each data set, the single variable representing the literacy plausible values has been renamed as PVLIT. If the SAS syntax is used to generate the data, we provide a simple SAS macro (%plausible) to simplify the process (see Appendix 1 in the online version of the journal). To create these data sets using the SAS syntax, the "%plausible" macro first should be run first, followed by the following command:

%plausible (origdata = sub, dataname = Dataset01, dnum = 1);

For the "origdata," either a temporary data set or a permanent data set with the specified library location (i.e., libname) is required. The "dataname" is used for naming the temporary data set to be created. Finally, "dnum" indicates the data set number, and in this case, the value 1 indicates the first plausible value. For the data sets generated using either the SAS or R syntax, an additional text file containing a list of all data set names is required in Mplus (see Figures 1 and 2).

5. Export the data sets as a set of files that are in Mplus-compatible format.

Prior to importing into Mplus, each of the 10 newly created data sets needs to be exported as a file that is in ASCII text format. For SAS users, either the built-in SAS point-and-click "Export data" command or the provided syntax for exporting the SAS data set as txt (.dat or .csv) files can be used. We recommend creating the first data set both with and without the variable names. Although the files to be used as input data to Mplus should not contain the variable names, information about the names and order of the variables in the data sets are necessary for Mplus to process the data. We provide an exemplar SAS PROC EXPORT command in Appendix 1 in the online version of the journal. This process can be simplified and automated for the advanced SAS users. However, we find it useful to view each export outcome and the corresponding log file. In the event of a data processing error, SAS will automatically generate an error message within the log file. Similarly, the R syntax in Appendix 2 in the online version of the journal will export the newly created data sets as comma-delimited text (i.e., .csv) files to the user's working directory.

Upon completion of Steps 1 through 5, the 10 sets of Mplus-compatible data sets, one data set with the variable names, and one text file with the list of data set names should be present in one folder (see the image in Figure 2). On a related note, these files may be copied to different folders or locations. However, keeping them in a single folder is good practice to avoid potential programming errors (e.g., incorrect directory path).

File Home Share View

PIAAC in Mplus

Search PIAAC in Mplus

| Name	Date modified	Type	Size
datalist01	4/20/2018 3:29 PM	Text Document	1 KB
Dataset01	11/25/2019 12:32 PM	Microsoft Excel Comma Separated Values File	7,080 KB
Dataset01WithNames	11/25/2019 12:32 PM	Microsoft Excel Comma Separated Values File	7,080 KB
Dataset02	11/25/2019 1:54 PM	Microsoft Excel Comma Separated Values File	7,079 KB
Dataset03	11/25/2019 1:54 PM	Microsoft Excel Comma Separated Values File	7,080 KB
Dataset04	11/25/2019 1:54 PM	Microsoft Excel Comma Separated Values File	7,080 KB
Dataset05	11/25/2019 1:54 PM	Microsoft Excel Comma Separated Values File	7,080 KB
Dataset06	11/25/2019 1:54 PM	Microsoft Excel Comma Separated Values File	7,080 KB
Dataset07	11/25/2019 1:54 PM	Microsoft Excel Comma Separated Values File	7,080 KB
Dataset08	11/25/2019 1:54 PM	Microsoft Excel Comma Separated Values File	7,079 KB
Dataset09	11/25/2019 1:54 PM	Microsoft Excel Comma Separated Values File	7,080 KB
Dataset10	11/25/2019 1:54 PM	Microsoft Excel Comma Separated Values File	7,080 KB

FIGURE 1. *Screenshot of the data folder.*
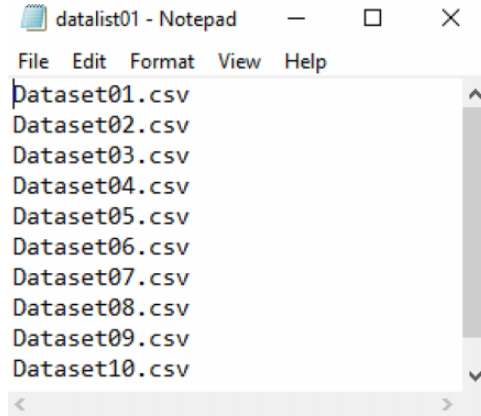
FIGURE 2. *An example of a text file containing the data file names for input to Mplus.*

Statistical Analysis

Example #1—Predicting motivation to learn from literacy skill.

To examine how motivation to learn was predicted by literacy skills, we constructed a structural equation model (Brown, 2014; Kline, 2016). The analysis was conducted in two steps. First, the measurement model was fitted using the four motivation to learn items as observed indicators of a single latent construct, specifying FIML estimation of point estimates, and using the paired jackknife (jackknife 2) method estimation to estimate standard errors (OECD, 2016). The CFA model was specified based on the theoretical proposition of Gorges et al. (2016). After fitting the measurement model, the structural model next was fitted to examine how the latent motivation to learn construct was predicted by literacy skills. Model building was conducted sequentially, starting with an unconditional model and terminating with the fully conditional model. The obtained parameter estimates obtained from Mplus represent the pooled estimates from analyses conducted on each of the 10 generated data sets, where each analysis uses a distinct set of plausible values. Both the measurement model and structural model were fitted using the supplied sampling weights (SPFWT0) and replicate weights (SPFWT1 to SPFWT80). However, because model fit indices are not generated by Mplus when replicate weights are used, these models additionally were fitted without the use of the replicate weights to obtain estimated model fit indices. Appendix 3 in the online version of the journal provides the Mplus syntax for fitting the structural model.

Example #2—Predicting literacy skill from demographic characteristics, socioeconomic status, and health status.

Using SEM, literacy skill was modeled as a function of selected demographic characteristics, socioeconomic status, and health status. The model is similar to linear regression, but, based on results from our preliminary analyses, two of the covariances between the predictor variables were explicitly constrained to zero (see Figure 3 and Appendix 4). As described in Example 1, we fitted the model using the replicate weights, then refitted the model without the use of the replicate weights to obtain fit indices.

To evaluate the models, we use the criteria provided by Kline (2016). Specifically, good model fit was indicated by the following values for fit indices: comparative fit index (CFI) > .90, root mean squared error of approximation (RMSEA) < .10, and standardized room mean squared residual (SRMR) < .10. The obtained estimate of R^2 also was inspected. When the replicate weights were applied, the jackknife 2 standard error estimation method was employed (OECD, 2016). The data preparation described previously was carried out using SAS Version 9.4 (SAS Institute Inc., 2002–2012) and R Version 1.1.423, and all structural equation models were fitted using Mplus Version 8 (Muthén & Muthén, 1998–2017).

Results

The weighted descriptive summary for each variable of interest is presented in Table 1.

Example #1—Predicting motivation to learn from literacy skill.

The model specification and results are summarized in Figure 4. The measurement model for the latent motivation to learn construct showed good model fit (CFI = .98, RMSEA = .08, SRMR = .02). Therefore, we concluded that the motivation to learn construct demonstrated good evidence of validity in this study. The structural model showed good fit (CFI = .96, RMSEA = .05, SRMR = .03) and additionally showed that literacy skill positively predicted motivation to learn ($b = 0.01$, $p < .05$). Note that the differences in the model fit indices obtained with the 10 generated data sets—each containing a different plausible value variable—were minor (within two decimal points).

Example #2—Predicting literacy skill from demographic characteristics, socioeconomic status, and health status.

The model specification and results are summarized in Figure 3. Model fit was adequate (CFI = .94, RMSEA = .05, SRMR = .03). As occurred with Example 1, differences in the model fit indices obtained with the 10 generated data sets were minor (within two decimal points). Age, educational attainment, self-rated health, and race/ethnicity significantly predicted literacy skill. For example, on average, adults with college or higher degree had literacy skill that was approximately 36 points higher ($b = 35.89$, $p < .001$) than those with less than a college degree. Additionally, the average literacy skill score of Hispanic adults was significantly lower than that of White adults by about 48 points ($b = -47.75$, $p < .001$).
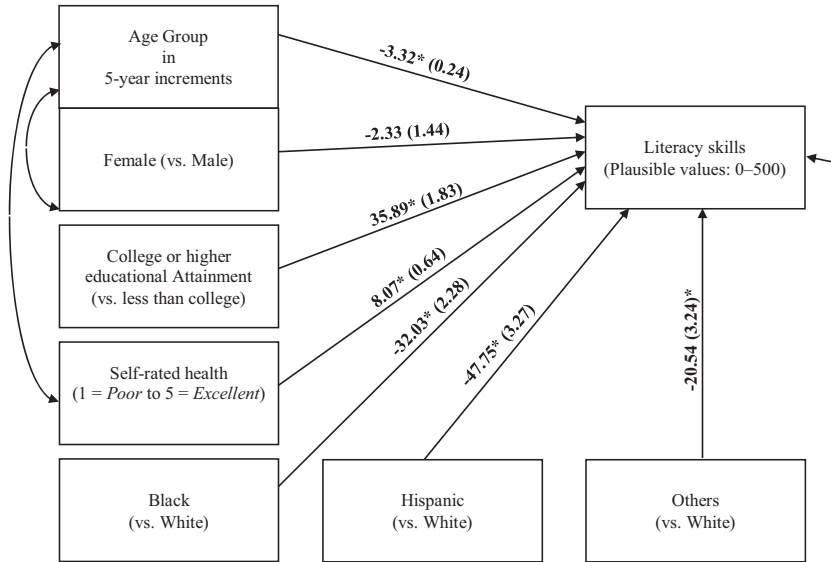
FIGURE 3. *(Example 1) Path diagram for structural equation model with estimated coefficients and standard errors. Note. Sampling weights have been applied, and standard errors have been estimated using the replicate weights. Model fit: $\chi^2(7) = 115.65$, $p < .01$; comparative fit index = .94; root mean squared error of approximation = .05; standardized room mean squared residual = .03; R^2 = .38. *$p < .05$.*

In our follow-up analysis to ensure that the analytic approach in Mplus was consistent with the established approach in the IDB analyzer, we examined several probit and linear regression models using the set of literacy plausible values for the outcome as well as the predictor. For example, we examined the associations between literacy, adult education, training participation (yes vs. no), and gender (women vs. men). Despite the fact that WLSMV was not an available estimation option in SAS, all estimated coefficients, standard errors, and *p* values were virtually identical (results are not reported here but are available upon request). Although SEMs cannot be estimated using the IDB analyzer/SAS, these follow-up analyses show preliminary validity of the analytic approach with the PIAAC plausible values and replicate weights in Mplus. It should be noted, however, that a regression model such as this that uses the estimated plausible values as independent (predictor variables) may not be consistent with how the provided plausible values in the PIAAC data were generated. Although the PIAAC technical report (OECD, 2013) provides for the use of the provided plausible values as either independent or dependent variables, as does the IDB Analyzer application, Schofield et al. (2015) demonstrate that the use of these plausible values as independent variables results in biased inferences and

TABLE 1.
Weighted Descriptive Summary

Variables	Mean (Standard Error) or Percentage
Literacy score (0–500 points)	269.28 (1.08)
Motivation to learn	
Item 1: "I like learning new things"	4.17 (0.02)
Item 2: "I like to get to the bottom of difficult things"	3.94 (0.02)
Item 3: "I like to figure out how different ideas fit together"	3.75 (0.02)
Item 4: "If I don't understand something, I look for additional information to make it clear"	4.12 (0.01)
Age-group	7.04 (0.02)
Gender (female)	51.90%
Race and ethnicity	
White	68.30%
Black	11.90%
Hispanic	12.70%
Others	7.10%
Educational attainment (College degree or higher)	40.30%
Self-rated health	3.51 (0.02)

Note. N = 6,632; motivation to learn items were coded from 1 = *not at all* to 5 = *to a very high extent*; self-rated health was coded from 1 = *poor* to 5 = *excellent*. Age-group was an ordinal variable coded (from 3 = *25–29 years* to 12 = *71+ years*); Program for the International Assessment of Adult Competencies sampling weights were applied, and replicate weights were applied using the paired jackknife (i.e., jackknife 2) method.

recommend an alternative approach that employs mixed effects structural equation modeling to more appropriately account for measurement error in these predictors. Such an approach requires access to the original item-level responses and, ideally, the original measurement model that was used to construct the latent proficiency scale.

Discussion and Conclusions

This article provided practical guidance in how to analyze large-scale assessment data with Mplus using the PIAAC data as an example. Use of Mplus can expand the scope of analysis and enable researchers to fit more complex statistical models. Data preparation may seem daunting, but, as can be seen in our example syntax for the data preparation and demonstration of simple analyses, it can be automated quite easily. We demonstrated this data preparation in several steps and also demonstrated simple inferential analyses in Mplus using the publicly available PIAAC data. Our sample syntax is useful particularly for researchers who are not yet familiar with Mplus and/or PIAAC data. Although our
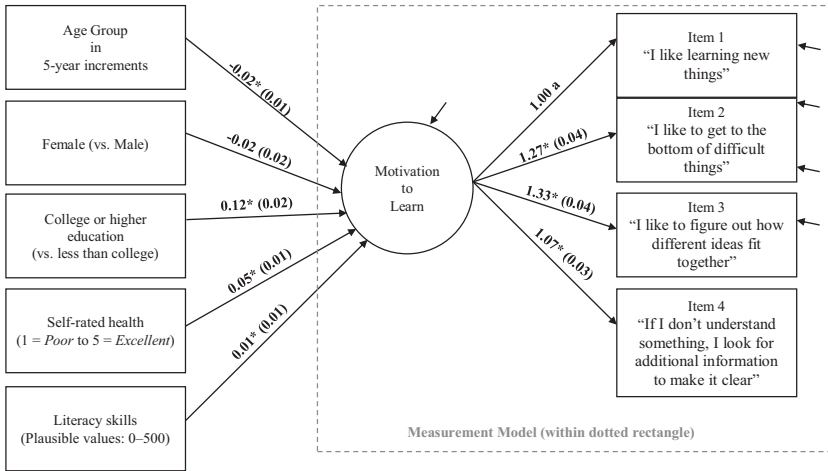
FIGURE 4. *(Example 2) Path diagram for structural equation model with estimated coefficients and standard errors. Note. "a" is fixed to 1. Sampling weights have been applied, and standard errors have been estimated using the replicate weights. Model fit for measurement model:* $\chi^2 = 74.70^*$, *comparative fit index (CFI) = .98, root mean squared error of approximation (RMSEA) = .08, standardized room mean squared residual (SRMR) = .02; model fit for structural model:* $\chi^2 = 292.62^*$, *CFI = .96, RMSEA = .05, SRMR = .03.* $*p < .05$.

suggested data preparation and analyses are just one of many ways to analyze large-scale assessment data, our intention is to provide practical guidance for education researchers. As stated earlier, resources such as Scherer (2020) provide an excellent resource for technical details about fitting other specific SEM models (e.g., multilevel SEMs). Additionally, although it is beyond the focus of the current study, alternatives to using the institutionally provided plausible values also exist, and these alternatives should be considered, for example, when the analyst desires to fit an analysis model that is more complex than the imputation model used to generate the institutionally provided plausible values. Generating such plausible values would be appropriate, for example, when using the PIAAC data (where the institutionally provided plausible values are generated from an IRT framework using a structural regression model) to fit a latent class model. Mplus provides such functionality to generate plausible values, provided the original item scores are available upon which the latent constructs are based (see, e.g., Asparouhov & Muthén, 2010; Carlin, 1992; Clark & Muthén, 2009). At the same time, there are several limitations to using applications such as Mplus for analysis of large-scale assessment data such as PIAAC. First, unlike the IDE application and the IDB Analyzer, data preparation can be time-consuming and may result in errors (e.g., programming errors, incorrect

designation of missing values). It is critical to repeatedly check both original and recoded data as well as data preparation accuracy at each step. Second, the use of the plausible values and replicate weights do not allow Mplus to produce the commonly used model fit indices. As stated earlier, the use of the replicate weights is recommended to estimate standard errors. Yet, to obtain fit indices, the model needs to be estimated without the use of replicate weights, although this approach to obtaining fit indices should be acknowledged as a methodological limitation. When using this approach, we recommend fitting the model separately with each distinct set of plausible values and verifying the consistency among the resultant fit indices.

Third, while not considered in our demonstration, available options for the use of link functions may be somewhat limited in Mplus. That is, when the outcome variable is categorical, certain types of SEM only allow the use of a probit link function, and it can be difficult to interpret results in these instances. Fourth, if a researcher is interested in comparing data from multiple countries, the complex sampling strategies employed as well as method of standard error estimation that is appropriate (e.g., jackknife vs. jackknife 2) may differ across the countries. Finally, although this article focused solely on the plausible values that are provided in the PIAAC data and existing functions in Mplus, more sophisticated approaches such as estimation of original skill proficiency models with latent variables are feasible, as outlined, e.g., in Schofield et al. (2015). As such, readers should be aware of other analytic approaches to PIAAC data and additional capabilities of Mplus. At the time of this study, few resolutions to these methodological limitations have been proposed, and, until consensus on approaches to addressing these limitations has been achieved, the researchers should explicitly acknowledge such limitations in their reports.

Despite its limitations, the capability of Mplus to fit a variety of models such as SEM, multilevel models, growth models, mixture models, and Bayesian models certainly can expand the scope of analysis for the large-scale assessment data. Correspondingly, as models posited and assessed in the literature become more sophisticated and their ability to explain educational phenomena is more nuanced, the need for researchers who are facile with the use of software applications for these data will increase. Also, increased interest in cross-national educational comparisons will drive a need for sophisticated models—multigroup SEMS, for example—that can effectively facilitate these comparisons in meaningful ways. Large-scale assessment data are collected with methodological sophistication yet are provided in a user-friendly format. Moreover, they offer the opportunity for the use of advanced statistical modeling applications, and the use of analytic tools that facilitate this should be more readily discussed and employed in education research communities.

Authors' Note

Declaration of Conflicting Interests

Funding

ORCID iD

Takashi Yamashita ⓘ https://orcid.org/0000-0003-2325-126X

References

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling* (Vol. 243–277, pp. 277). Lawrence Erlbaum.

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(3), 359–388. https://doi.org/10.1080/10705511.2017.1406803

Asparouhov, T., & Muthén, B. (2010). *Plausible values for latent variables using Mplus*. http://www.statmodel.com/download/Plausible.pdf

Avvisati, F., & Keslair, F. (2019). *REPEST: Stata module to run estimations with weighted replicate samples and plausible values*. https://EconPapers.repec.org/RePEc:boc:bocode:s457918

Brown, T. A. (2014). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Publications.

Carlin, J. B. (1992). Meta-analysis for 2×2 tables: A Bayesian approach. *Statistics in Medicine*, *11*(2), 141–158. https://doi.org/10.1002/sim.4780110202

Caro, D. H., & Biecek, P. (2017). intsvy: An R package for analyzing international large-scale assessment data. *Journal of Statistical Software*, *81*(7), 1–44. https://doi.org/10.18637/jss.v081.i07

Clark, S. L., & Muthén, B. (2009). Relating latent class analysis results to variables not included in the analysis. https://www.statmodel.com/download/relatinglca.pdf

Cresswell, J., Shwantner, U., & Waters, C. (2015). *A review of international large-scale assessments in education: Assessing component skills and collecting contextual data*. https://www.oecd.org/fr/developpement/a-review-of-international-large-scale-assessments-9789264248373-en.htm

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, *22*(2), 153–164.

Gorges, J., Maehler, D. B., Koch, T., & Offerhaus, J. (2016). Who likes to learn new things: Measuring adult motivation to learn with PIAAC data from 21 countries. *Large-Scale Assessments in Education*, *4*(1), 9. https://doi.org/10.1186/s40536-016-0024-4

IBM Corp. (2016). *IBM SPSS Statistics for Windows* (Version 24.0). https://www.ibm.com/analytics/spss-statistics-software

International Association for the Evaluation of Educational Achievement. (2016). *Help manual for the IDB analyzer (SAS macros)*. http://www.iea.nl/data

Kirsch, I., & Lennon, M. L. (2017). PIAAC: A new design for a new era. *Large-Scale Assessments in Education*, *5*(1), 11. https://doi.org/10.1186/s40536-017-0046-6

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). The Guilford Press.

McFarland, J., Hussar, B., Wang, X., Zhang, J., Wang, K., Rathbun, A., Barmer, A., & Bullock Mann, F. (2018). *The condition of education 2018*. https://nces.ed.gov/pubs2018/2018144.pdf

Murray, T. S., Kirsch, I. S., & Jenkins, L. B. (1998). *Adult literacy in OECD countries: Technical report for the first International Adult Literacy Survey* (NCES98-053, Issue). https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=98053

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus statistical analysis with latent variables user's guide*. Muthén & Muthén.

National Center for Education Statistics. (n.d.). *History of international adult literacy assessments*. https://nces.ed.gov/surveys/piaac/history.asp

OECD. (2016). *Technical report of the Survey of Adult Skills (PIAAC)*. OECD Publishing. http://www.oecd.org/skills/piaac/PIAAC_Technical_Report_2nd_Edition_Full_Report.pdf

Organization for Economic Cooperation and Development. (2009). *PISA data analysis manual*. https://www.oecd.org/pisa/pisaproducts/pisadataanalysismanualspssandsassecondedition.htm

Organization for Economic Cooperation and Development. (2016). *Technical report of the Survey of Adult Skills (PIAAC)*. OECD Publishing. http://www.oecd.org/skills/piaac/PIAAC_Technical_Report_2nd_Edition_Full_Report.pdf

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, *39*(2), 142–151. https://doi.org/10.3102/0013189x10363170

SAS Institute Inc. (2002–2012). *SAS* (Version 9.4). https://support.sas.com/documentation/94/

Scherer, R. (2020). Analysing PIAAC data with structural equation modelling in Mplus. In D. B. Maehler & B. Rammstebt (Eds.), *Large-scale cognitive assessment* (pp. 165–208). Springer. https://library.oapen.org/bitstream/handle/20.500.12657/41286/2020_Book_Large-ScaleCognitiveAssessment.pdf?sequence=1#page=170

Schofield, L. S., Junker, B., Taylor, L. J., & Black, D. A. (2015). Predictive inference using latent variables with covariates. *Psychometrika*, *80*(3), 727–747. https://doi.org/10.1007/s11336-014-9415-z

Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using Mplus*. John Wiley.

Authors

TAKASHI YAMASHITA is an associate professor at the Department of Sociology, Anthropology, and Health Administration and Policy, University of Maryland, Baltimore County, Public Policy Building, PUP252, 1000 Hilltop Circle, Baltimore, MD 21250, USA; email: yamataka@umbc.edu. His research interests are social determinants of health and well-being over the life course; wider benefits of lifelong and lifewide learning; health literacy; quantitative research methods (statistical methods, information visualization); gerontology education; and social statistics education.

THOMAS, J. SMITH is a professor at the Department of Educational Technology, Research and Assessment, Northern Illinois University, Gabel Hall, DeKalb, IL 60115, USA; email: tjsmith@niu.edu. His research interests are Statistics and research methodology, and analysis of large-scale national/international data.

PHYLLIS A. CUMMINS is a senior research scholar at The Scripps Gerontology Center, Miami University, Upham Hall, 100 Bishop Circle, Oxford, OH 45056, USA; email: cumminpa@miamioh.edu. Her research interests are education and training for middle-aged and older workers, the role community colleges play in education and training for older adults, and skill development over the life course.