NEPS
National Educational Panel Study

Survey Papers of the German National Educational Panel Study (NEPS)
at the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg

The NEPS Survey Paper Series provides articles with a focus on methodological aspects and data handling issues related to the German National Educational Panel Study (NEPS).

The NEPS Survey Papers are edited by a review board consisting of the scientific management of LIfBi and NEPS.

They are of particular relevance for the analysis of NEPS data as they describe data editing and data collection procedures as well as instruments or tests used in the NEPS survey. Papers that appear in this series fall into the category of 'grey literature' and may also appear elsewhere.

The NEPS Survey Papers are available at https://www.neps-data.de (see section “Publications“).

Editor-in-Chief: Corinna Kleinert, LIfBi/University of Bamberg/IAB Nuremberg

Contact: German National Educational Panel Study (NEPS) – Leibniz Institute for Educational Trajectories – Wilhelmsplatz 3 – 96047 Bamberg – Germany – contact@lifbi.de

Estimating Plausible Values with NEPS Data: An Example Using Reading Competence in Starting Cohort 6

Anna Scharl, Leibniz Institute for Educational Trajectories
Claus H. Carstensen, University of Bamberg
Timo Gnambs, Leibniz Institute for Educational Trajectories

Preprint

E-mail address of lead author:

anna.scharl@lifbi.de

Bibliographic data:

Scharl, A., Carstensen, C. H., & Gnambs, T. (2020). *Estimating Plausible Values with NEPS Data: An Example Using Reading Competence in Starting Cohort 6* (NEPS Survey Paper No. ∞). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Estimating Plausible Values with NEPS Data: An Example Using Reading Competence in Starting Cohort 6

Abstract

The National Educational Panel Study (NEPS) provides data on the development of various competence domains across the life span. Because research questions using these competences typically pertain to latent relationships between constructs, this paper gives an overview of the concept of plausible values and how to estimate unbiased effects that account for measurement error in competence scores. Plausible values incorporate responses to a competence test as well as various background variables. Only if all variables relevant for the specific research question are part of the background model, plausible values estimate unbiased population-level effects. Because the NEPS allows for a multitude of different research questions and, by design, provides a large and growing amount of background information, it is difficult to provide plausible values that fit each conceivable research question. Therefore, the R routine `plausible_values()` in the package **NEPS*scaling*** was developed. Its functionality enables NEPS data users to easily generate custom-tailored plausible values addressing their specific research needs. Because missing data are a pervasive problem in large-scale assessments, **NEPS*scaling*** also offers a sequential Classification and Regression Trees (CART) algorithm for handling missing values in background variables. This paper introduces the concept of plausible values and CART. Moreover, an applied example demonstrates how to estimate plausible values with `plausible_values()`.

Keywords
plausible values, missing values, classification and regression trees, multiple imputation, competences

Contents

1 Introduction

The National Educational Panel Study (NEPS) provides data on educational trajectories of participants in the German educational system from birth to retirement (Blossfeld et al., 2011). As such, NEPS data can be used to investigate diverse research questions that, among others, might address important antecedents, the development, or potential returns of domain-specific competencies. The unique design of the NEPS allows an integrated perspective that includes personal characteristics such as gender or the socio-economic background as well as context variables in the form of, for example, school or workplace characteristics. In the NEPS, several competence domains are assessed including (among others) mathematics, reading, science, and information and communication technology literacy (Fuß et al., 2019; Weinert et al., 2011). These are typically scaled using models of item response theory (for a detailed description of the scaling procedure see Pohl and Carstensen, 2012) and published as weighted maximum likelihood estimates (WLEs; Warm, 1989) in the Scientific Use Files (SUFs).

Although WLEs give an accurate representation of an individual's competence level, they systematically overestimate the variance in a sample and lead to underestimated correlations and regression coefficients on a population level (Lüdtke & Robitzsch, 2017; von Davier et al., 2009). One reason for this bias are ignored effects of third variables on the competence level, such as gender, socio-economic status, educational level, or general cognitive abilities (Lüdtke & Robitzsch, 2017; von Davier et al., 2009; Wu, 2005). In contrast, those background variables (also called conditioning variables) are taken into account as latent regressors (see Figure 1) in the estimation of plausible values (Mislevy, 1991). As a result, plausible values allow for unbiased estimates of population effects although they are no longer unbiased scores for individual respondents. Thus, as long as group-level effects are the focus of interest (as is typical in scientific research) plausible values are superior to point estimates such as WLEs and allow for more precise estimates of true effects. For the estimation of plausible values, typically, the background variables have to be fully observed. But despite best efforts, non-response cannot be completely eliminated in large-scale assessments such as the NEPS (Zinn & Gnambs, 2018). The estimation of plausible values, therefore, has to incorporate an appropriate method of dealing with missing information.

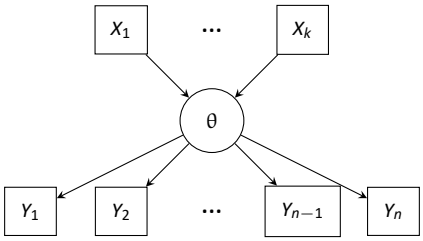

Figure 1: Latent regression model for estimating plausible values for the true competence θ *with k covariate background variables* X_1 *to* X_k *and n indicator variables* Y_1 *to* Y_n *for the measured competence.*

In other large scale assessment studies (LSAs) such as the Programme for International Student Assessment (PISA; OECD, 2017) or the Trends in International Mathematics and Science Study (TIMSS; Martin et al., 2016), plausible values are provided in the SUFs. Because the NEPS provides a large and, by design, growing amount of data over the life span and, thus, allows for a

broad variety of research questions involving competence scores, it is rather difficult to specify a single background model for each starting cohort that fits each conceivable research question. Therefore, we provide the R function `plausible_values()` in the package **NEPS*scaling*** for NEPS data users to estimate plausible values tailored to the specific research question.

In the following sections, we give an overview of the plausible values technique and introduce sequential classification and regression trees (CART; Burgette & Reiter, 2010) as our choice of handling missing data in the background data. Then, the package **NEPS*scaling*** and the function `plausible_values()` are introduced by a practical example using reading competencies in starting cohort 6 (adults). A step by step user guide for the package shows how to generate plausible values and export them into formats to be used by other statistical software such as SPSS (IBM Corp, 2015), Stata (StataCorp, 2015), or Mplus (Muthén & Muthén, 1998-2017).

2 Estimating plausible values

The plausible values technique can be considered a special case of multiple imputation of missing data (Little & Rubin, 1987; Rubin, 1987). While imputed variables are partially observed, competences are latent constructs and, thus, are never directly observed. Therefore, competences are inferred from a number of observed indicators (i.e., responses to the items of a competence test) that are afflicted with measurement error. Moreover, if associations between competencies and third variables (e.g., gender) are of interest, these third variables need to be taken into account when estimating the latent construct. Incorporating covariates of the latent ability into the estimation of latent competences accounts for measurement error arising from ignoring structures and relationships in the data that influence the competence (see Figure 1). For example, it is well known that male students typically achieve higher scores on mathematical tests than female students (e.g., OECD, 2015). Thus, if gender is not taken into account for the estimation of mathematical competence, competence estimates on the population level would be biased (although competence scores for individual students would be unbiased). This highlights that plausible values must not be used for individual feedback for specific students because the mean of the ability distribution is shifted by incorporating additional information. However, plausible values result in more precise effects on the population level. A closer look on how plausible values are estimated makes this point more straightforward:

Plausible values are random draws from a posterior ability distribution $p(\theta_i|\vec{y}_i)$ for the true competence θ of subject *i* given the response vector $\vec{y}_i$ for the items of the competence test:

$$p(\theta_i|\vec{y}_i) \propto p(\vec{y}_i|\theta_i) \cdot p(\theta_i) \tag{1}$$

Here, θ_i is the true competence of subject *i*, $p(\vec{y}_i|\theta_i)$ is the item response model, that is, a Rasch (1960) model in the NEPS (see Pohl & Carstensen, 2012), and $p(\theta_i)$ is the population model. At first, we assume that the competence is normally distributed in the population:

$$p(\theta_i) \sim N(\mu, \sigma^2) \tag{2}$$

with μ and σ^2 being the population mean and variance. As the above example with mathematical competence and gender shows, there may be subgroups in the population. The competence is thus regressed on *k* covariates $\vec{x}_i = (x_{i1}, \ldots, x_{iK})$ in question (e.g., gender).

$$\theta_i = \beta_0 + \vec{x}_i\beta_K + \varepsilon_i \tag{3}$$

where β_0 and $\beta_K = (\beta_1, \ldots, \beta_K)^T$ are the regression coefficients and ε_i represents the residual. The regression parameters are now used to adapt the population model:

$$p(\theta_i|\vec{x}_i) \sim N(\beta_0 + \vec{x}_i\beta_K; \sigma^2_{\theta|\vec{x}_i}) \tag{4}$$

which leads to the posterior distribution used to draw multiple plausible values for subject *i*:

$$p(\theta_i|\vec{x}_i, \vec{y}_i) \propto p(\vec{y}_i|\theta_i) \cdot p(\theta_i|\vec{x}_i) \tag{5}$$

The mean of the distribution now changes depending on group memberships or, more general, depending on the variables in the background model and their respective values. The variance also reflects group variances instead of the population variance as a whole. This example can be extended to any number of covariates that are added in the regression as linear combinations (Wu, 2005). The challenge for applied researchers is the specification of a correct conditioning model for the research question at hand. Generally, *all variables that are part of the final analysis model* with respect to the latent trait should be included in the conditioning model. This also includes interactions or non-linear relationships.

It is possible to misspecify the conditioning model in two ways: by adding more variables than in the analysis model or by ignoring important variables. The first case is usually not severe, but, as long as the model stays identified in the face of very large amounts of background data, might even increase the precision of the estimated effects because additional information is included in the estimation of the latent trait (Lüdtke & Robitzsch, 2017; Meng, 1994). In contrast, neglecting to include important variables in the background model leads to biased outcomes on the population level (Bondarenko & Raghunathan, 2016; Lüdtke & Robitzsch, 2017; Meng, 1994). Therefore, researchers need to carefully decide which variables to include in the background model.

As each subject receives a set of random draws from the distribution with density $p(\theta_i|\vec{x}_i, \vec{y}_i)$, the competence scores vary within each subject. These variations reflect the uncertainty of the estimation process (Lüdtke & Robitzsch, 2017; von Davier et al., 2009; Wu, 2005). As a consequence, empirical analyses need to incorporate multiple plausible values. For a long time, it has been suggested that 5 plausible values might suffice. However, recent analyses suggested that more plausible values (e.g., at least 20 or 30) are preferable (see Bodner, 2008; Graham et al., 2007). Further information on how to analyse plausible values is described in Little and Rubin (1987, p. 257), Mislevy (1991, p. 182), and Rubin (1987, p. 76f.). Also, more detailed and comprehensive summaries of the plausible values technique in general can be found in Lüdtke and Robitzsch (2017), von Davier et al. (2009), and Wu (2005).

A challenge in the estimation of plausible values is missing data in the conditioning variables. In LSAs, it is virtually impossible to fully observe all variables for all respondents because some participants will refuse to provide responses to selected items (Zinn & Gnambs, 2018). Therefore, different ways of handling missing information for the estimation of plausible values have been devised. In the following, we will shortly discuss the usually used method of missing in-

dicators (Martin et al., 2016; Martin et al., 2007; OECD, 2017) and its disadvantages as well as our approach to solving them.

2.1 Missing indicators

A common way of obtaining plausible values with missing values on the background variables follows a two-step procedure. First, all background variables are appropriately recoded. Nominal or ordinal responses are dummy coded (Martin et al., 2016; Martin et al., 2007; OECD, 2017) and metric responses are criterion scaled (Beaton, 1969), for example, as in Martin et al. (2007). Then, a principle component analysis reduces the number of predictors in the background model to those components explaining most of the variance (e.g., 90%; OECD, 2017). The principle components, and possibly some important primary variables such as gender, socio-economic background and other context variables, are then used in the background model to estimate the plausible values (Martin et al., 2016; OECD, 2017).

This coding strategy allows for the integration of missing values information via dummy indicator variables or by defining persons with missing values as a distinct group in criterion scaling. Although this approach is widely used in LSAs such as PISA or the Programme for the International Assessment of Adult Competencies (PIAAC; OECD, 2017; Yamamoto et al., 2013), creating a dummy indicator variable for missing data has been criticized by some authors as merely redefining the model (Schafer & Graham, 2002) and not taking the dependencies between the missing values and latent ability into account (Aßmann et al., 2015). Furthermore, it has been shown that missing indicator methods can lead to biased regression coefficients or residual variances (Jones, 1996) and even a mean shift of the ability estimates (Rutkowski, 2011) that could lead to severely biased results when comparing populations of which, for example, only one exhibits biased means due to missingness. Therefore, we decided to use sequential classification and regression trees (CART; Burgette & Reiter, 2010; Doove et al., 2014; Loh, 2011) to impute the missing values. The technique and its integration into the plausible values estimation procedure are discussed in the following sections.

2.2 Multiple imputation via sequential CART

Classification and regression trees (CART) were introduced by Breiman et al. (1983). The CART algorithm is used to predict a subject's value on one variable given the subject's values on a range of predictor variables (Burgette & Reiter, 2010; Doove et al., 2014; Loh, 2011). To achieve this prediction, the algorithm recursively splits the variable space of the outcome variable into binary partitions until a node purity criterion is met. A random draw of the values in the final partitions is then used as the predictions for the missing values. The partitioning can be represented by a tree structure with the partitions as tree nodes and the partitioning decisions as edges (see Figure 2). The final partitions are the tree's leaves. Depending on the measurement level of the outcome variable a classification (nominal or ordinal level) or a regression tree (metric level) is constructed. They differ in their node purity criteria.

Consider the example of an ordinal outcome variable X_0 with values $x \in \{1, 2, 3\}$ and the predictor variables X_1 (metric) and X_2 (binary) depicted in Figure 2. Depending on the value of X_2, X_1 has a different relation to the outcome variable X_0. For $X_2 = 0$, the cases with $X_1 \leq -1$ are classified as 2, otherwise as 3, whereas for $X_2 = 1$, the cases with $X_1 \leq -0.3$ are classified as

2, otherwise as 1. The tree thus models complex dependencies and interaction effects without explicitly specifying them (Burgette & Reiter, 2010; Doove et al., 2014).

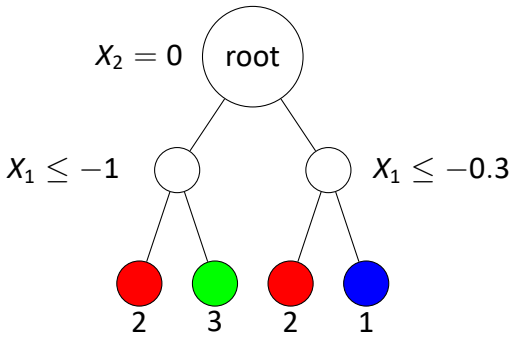

Figure 2: Example of a classification tree for an outcome variable with three classes. Only if the condition is satisfied, a case goes to the left child node (adapted from Loh, 2011, p. 15)

The CART algorithm is embedded in a Markov Chain Monte Carlo procedure. Let X be an $n \times k$ matrix with n observations on k variables. The columns of X contain missing values, i.e., the complete $X_{com,j}$, $j = (1, \ldots, k)$, is composed of the observed values $X_{obs,j}$ and the missing values $X_{mis,j}$. CART then imputes the missing values by following five steps:

Step 1: Substitute the missing values X_{mis} with initial values (e.g., means) for all variables in X with missing portions.

Step 2: For all X_j with missing portions, construct trees using all other variables X_{-j} as predictors.

Step 3: For all X_j with missing portions, replace $X_{mis,j}$ with predictions from the trees.

Step 4: Repeat steps 2-3 i times with a burnin period b to ensure convergence to a stationary data distribution.

Step 5: Randomly draw m data sets from the $i - b$ sampled data sets of the data distribution.

The variables are ordered to have increasing proportions of missing values. Thus, the trees are first constructed for the variables with the most observed observations. Those first predictions are subsequently used to construct trees for variables with more missing information.

2.3 Maximum likelihood estimation with nested multiple imputation

Adhering to the scaling guidelines set in Pohl and Carstensen (2012) a partial credit model (PCM; Masters, 1982) is estimated to accommodate binary as well as polytomous test items. The probability to respond in category y to item j is given by:

$$P(Y_{ij} = y|\theta, \delta) = \frac{exp\left\{\sum_{k=0}^{y}(\theta_i - \delta_{jk})\right\}}{\sum_{h=0}^{K_j} exp\left\{\sum_{k=0}^{h}(\theta_i - \delta_{jk})\right\}} \text{ with } \sum_{k=0}^{0}(\theta_i - \delta_{jk}) \equiv 0 \qquad (6)$$

where Y_{ij} denotes the response of person i on item j, K_j denotes the total number of categories for item j, θ_i denotes the latent competence of person i, and δ_{jk} the item parameters. To counter

possible larger impact of polytomous items on the estimated parameters, they are weighted by 0.5 (Pohl & Carstensen, 2012). For binary items Equation 6 reduces to the well-known Rasch (1960) model:

$$P(Y_{ij} = 1|\theta, \xi) = \frac{exp\{\theta_i - \xi_j\}}{1 + exp\{\theta_i - \xi_j\}} \tag{7}$$

with item difficulty parameters ξ_j.

In a first step, the item response model is fit to the competence test data. In this step, item parameters and individual likelihoods are estimated using the Expectation-Maximization algorithm implemented in the R package **TAM** (Robitzsch, Kiefer, et al., 2017). In a second step, plausible values are drawn for each person, also using **TAM**.

If longitudinal plausible values are to be estimated, a PCM is fit to each measurement time point separately. The plausible values of later measurement points are then transformed to incorporate link information obtained in the scaling procedure (cf. Fischer et al., 2016). This approach was chosen to exactly mirror the procedures established in the NEPS.

To ensure complete background variables, we apply the concept of nested multiple imputation as introduced by Weirich et al. (2014): before plausible values are estimated, the missing values in the conditioning variables are imputed by the CART algorithm as described in the previous section.

The multiply imputed data sets are each used to estimate a range of plausible values for the ability. This results in *number of multiple imputations* $\times$ *number of plausible values* competence estimates for each subject. To keep the output of **NEPS*scaling*** concise, only the pre-specified number of plausible values is returned as a random subset of all estimated plausible values. Nested multiple imputation can consider dependencies between the ability and the conditioning variables if an ability indicator is included in the imputation model (Weirich et al., 2014). Still, only a proxy for the missing ability is used (e.g., the WLEs provided in the SUFs) that is itself not conditioned on the background variables.

Again, the importance of a correctly specified conditioning model needs to be stressed. It is therefore recommended to not only include predictor variables with regard to the latent trait, but also to include predictors for the background variables themselves to increase the precision of the estimation and avoid model misspecifications.

3 Estimating plausible values for NEPS data in R

The R package **NEPS*scaling*** provides users of NEPS data with a way of estimating plausible values for the major competence domains. The estimation by `plausible_values()` is based on the psychometric results described in the respective technical reports of the substudies. To further ensure comparability between the plausible values and the WLEs, any corrections of the WLEs (e.g., for sample dropout, changes in the booklet rotation design, or linking) are acknowledged by the function (see the respective technical reports for potential corrections

applied).

For users unfamiliar with R a number of books and online tutorials[1] are available that give a gentle and yet comprehensive introduction into the basics of R (e.g., Field et al., 2012). Moreover, we recommend using a development environment like RStudio (http://www.rstudio.com) to run the syntax demonstrated below. Because an introduction into R is beyond the scope of this paper, we recommend familiarizing with R before continuing. Thus, in the following, we assume basic knowledge of the R language such as assigning objects or addressing variables in a `data.frame`.

3.1 Installation of NEPS*scaling*

Before installing **NEPS*scaling***, please note that **NEPS*scaling*** does not contain any rawdata from the SUFs. Users have to apply for access to the NEPS data in line with the procedures outlined on the NEPS website (http://www.neps-data.de) and download the SUFs to their local computer.

To use **NEPS*scaling***, the package has to be installed in R using the standard method. The URL of the package can be looked up at https://www.neps-data.de/PV.

```
install.packages("[INSERT URL HERE]", repos = NULL, type = "source")
```

The package depends, among others, on the R packages **haven** (Wickham & Miller, 2016), **dplyr** (Wickham et al., 2017), and **TAM** (Robitzsch, Kiefer, et al., 2017). These dependencies are automatically installed during the installation of **NEPS*scaling***.

3.2 Steps to estimate plausible values

After the installation of **NEPS*scaling*** is complete, the estimation and analysis of plausible values follows six steps:

1. Download the NEPS SUFs and extract them into a dedicated directory on the local computer. This directory needs to be accessible by `plausible_values()`; the data must be either SPSS (`.sav`) or Stata (`.dta`) files (the default file formats for NEPS SUFs).
2. Prepare the data for your background model. It is not required to do this in R; any software can be used.
3. Load the package **NEPS*scaling*** in R, and import the prepared data for your background model.
4. Estimate plausible values for the specified starting cohort, competence domain, and assessment wave with the function `plausible_values()` .
5. Save the generated plausible values.
6. Analyze the plausible values with the statistical program of your choice (you do not need to use R).

The function `plausible_values()` imports the NEPS competence file from the dedicated directory with the downloaded SUFs. It then processes the data so that an item response model that adheres to the NEPS scaling guidelines (Pohl & Carstensen, 2012) can be fit. Additionally, it checks the provided background data for compatibility and dummy-codes factor variables.

[1] e.g., http://www.cookbook-r.com/, or http://tryr.codeschool.com/

Subsequently, missing data in the background variables are imputed using CART and plausible values are estimated using marginal maximum likelihood estimation.

3.3 The function `plausible_values()`

The function `plausible_values()` in the package **NEPS*scaling*** takes the general form:

```

plausible_values(
    SC,
    wave,
    path,
    domain = c('MA','RE','SC','IC','LI','EF','NR','NT','OR','ST','BA','CD', 'GR'),
    bgdata = NULL,
    npv = 10L,
    longitudinal = FALSE,
    rotation = TRUE,
    min_valid = 3L,
    include_nr = TRUE,
    verbose = TRUE,
    control = list(EAP = FALSE, WLE = FALSE,
        ML = list(nmi = 10L, ntheta = 2000,
                  normal.approx = FALSE, samp.regr = FALSE,
                  theta.model = FALSE, np.adj = 8, na.grid = 5,
                  itermcmc = 100, burnin = 50, thin = 1,
                  cartctrl1 = 5, cartctrl2 = 0.0001))
)

```

The function `plausible_values()` accepts the following arguments:

Table 1: Description of the arguments to the function `plausible_values()`

Argument	Description and examples
`SC`	The starting cohort for which plausible values are to be estimated. • Starting cohort 5: `5` • Starting cohort 6: `6`
`wave`	The wave in which the competence test has been administered (see Fuß et al., 2019, for the respective wave numbers). • Wave 3: `3` • Wave 5: `5`
`path`	Refers to the file path where the competence data is stored on the local computer. • `'home/NAME/NEPS-data/'` • `'C:/Users/NAME/Documents/NEPS-data/'`
`domain`	The competence domain for which plausible values are to be estimated. The abbreviations follow the NEPS variable naming conventions (see Fuß et al., 2019). • Reading: `'RE'` • Science: `'SC'`

Argument	Description and examples
`bgdata`	The background data used in the conditioning model of the plausible values; it has to meet the following requirements: • it must be provided in the form of a `data.frame` in wide format • it has to contain the target identifier (`ID_t`) – it is recommended to use all available test takers per domain and wave for the estimation of the plausible values to use the full information in NEPS data and avoid unnecessary uncertainty in the estimation due to smaller sample sizes • categorical or ordered variables have to be formatted as `factors` If no background data is supplied (`bgdata = NULL`), plausible values are estimated for an empty population model.
`npv`	The number of plausible values to be returned. The default number is 10 (cf. Martin et al., 2016; Martin et al., 2007; OECD, 2017). However, more plausible values might lead to more precise results. • `npv = 20`
`longitudinal`	A logical with default FALSE. TRUE indicates that the competence scores are to be used for longitudinal research and, thus, models for multiple time points are estimated. The plausible values of longitudinal measurement points are transformed to incorporate the previously obtained link information (cf. Fischer et al., 2016). • `longitudinal = TRUE`
`rotation`	A logical with default TRUE. It indicates whether corrections for the position of the test in the test battery (1st or 2nd) should be applied (see respective technical reports for the competence tests, e.g., Koller et al., 2014). If `longitudinal` is set to TRUE, `rotation` is automatically set to FALSE (cf. Pohl & Carstensen, 2012). In rare cases, missing values occur on the rotation variable. These are coded as a separate rotation group. Because of this group's small size, not all possible response values are available for this group. The algorithms issue messages to alert to this circumstance. The messages can be ignored. • `rotation = FALSE`
`min_valid`	The minimum number of valid responses to the competence test for a test taker to be included in the plausible values estimation (defaults to 3; see Pohl & Carstensen, 2012). If `min_valid` is set to zero, all subjects listed in `bgdata` will receive plausible values. However, it is not recommended to estimate plausible values for subjects without any responses on the competence test, unless a large number of background variables are included that can be used as an imputation model. • `min_valid = 3`

Argument	Description and examples
`include_nr`	A logical with default TRUE. It indicates whether the number of not-reached items in the competence test should automatically be included in the background model as a proxy for processing speed. • `include_nr = FALSE`
`verbose`	A logical with default TRUE. It indicates whether the program's progress should be displayed in the console. • `verbose = FALSE`
`control`	A list of additional options. Next to more specific lists for the estimation algorithm and the CART algorithm, the logical parameters `EAP` and `WLE` controls if EAPs or WLEs are also estimated and returned. The logicals are set to FALSE by default. • `control = list(EAP = FALSE)`
`$ML`	A list of additional options for the estimation of plausible values and the imputation of missing data in the background model. Among others the number of multiple imputations (`nmi`) can be set here. The default number is 10. But the number should be adapted according to the amount of missingness in the data (see Graham et al., 2007). Use `?TAM::tam.pv` for further description of the possible arguments for the estimation of PVs. Furthermore, the list also contains the controls of the `CART` imputation algorithm. The default is 100 iterations with a 50 iterations burnin period. Note that the number of imputations cannot exceed `itermcmc - burnin`. The arguments have to be given in the form of a `list(arg1 = ..., )` • `control = list(ML = list(nmi = 20))`

The function returns an object of class `pv_obj`. It contains a record of all the arguments given to `plausible_values()` (e.g., starting cohort, wave, competence domain) and, depending on the arguments passed to the function `plausible_values()`, it returns several of the following values:

Table 2: Description of the return values of the function `plausible_values()`.

Object	Description
`pv`	The list of `data.frames` in which each `data.frame` contains one plausible value and the complete data used to estimate the plausible values.
`valid_responses_per_person`	The `data.frame` contains the target ID and the number of valid responses a test taker has on the competence test.
`EAP_rel`	EAP reliability. For each imputation, one value is returned. In the longitudinal case, the imputations are elements in a list; one reliability value per assessment time point is returned per element.
`eap`	The matrix contains the EAP and SE for each subject. Returned if `control = list(EAP = TRUE)`.
`WLE_rel`	WLE reliability for each processed assessment wave. Returned if `control = list(WLE = TRUE)`.
`wle`	The matrix contains the WLE and SE for each subject. Returned if `control = list(WLE = TRUE)`.
`type`	Whether `'cross'`-sectional or `'long'`-itudinal models have been estimated.
`position`	Indicator for testlet position. The position of the respective competence test can be read out using `attributes(get_test_position(pv_obj)$position)`. Returned if `longitudinal = FALSE` and `rotation = TRUE`.
`mean_PV`	Estimated latent mean of the plausible values (in the longitudinal case before transformation).
`regr_coeff`	Estimated latent regression coefficients of the background variables. In the cross-sectional case, a matrix with two columns per multiply imputed data set is returned. In the longitudinal case, a list of matrices is returned. Each matrix belongs to an imputed data set and its columns refer to the assessed time points.
`items`	Fixed item parameters and estimated standard errors.

Notes. Additionally, all arguments passed to the `plausible_values()` function initially are returned.

Moreover, various functions for easy access to the elements of the output object of class `pv_obj` are provided (e.g., `get_pv_index(pv_obj, index)` which returns the `data.frame` at position `index` in the list `pv`) as well as the function `write_pv()` to export the plausible values from R to the statistical programs SPSS, Stata (Version 14), or Mplus. `write_pv()` takes the arguments: `pv_obj` returned from `plausible_values()`, `path` specifying the location the object is to be exported to, and `ext` which denotes the intended program.

Furthermore, the parameter `min_valid` needs more consideration: The *minimum number of valid responses to a competence test* that is required to estimate plausible values for a test taker

is set to three by default (see the NEPS scaling standards, Pohl & Carstensen, 2012). But it is possible to estimate plausible values even for people who have not completed a test at all. If the background model is comprehensive enough it might be valid to estimate plausible values even if no information on a competence test is available. However, this requires careful model selection and a sound theoretical foundation. Generally, it is *not recommended to change* the default setting.

In general, maximum likelihood estimation is quite fast. Essentially, the computation times greatly depend on two variables. Firstly, the larger the number of models (i.e., whether cross-sectional or longitudinal plausible values are requested), the longer the computation takes. Secondly, the amount of background information (i.e., number of variables and proportion of missing values) determines the runtime of the CART algorithm. More variables and, especially, a higher proportion of missing values slow the algorithm down.

4 Example: Regressing reading comprehension on reading activities

To illustrate the estimation and analysis of plausible values with NEPS data, we replicate a study belonging to the dissertation project of Bonerad (2012). She investigated the impact of reading activities in leisure time and during working hours on reading competence. Additionally, several covariate variables (e.g., socio-economic status [SES] and age) were included in a structural equation model (SEM)[2]. The model is depicted in Figure 3.

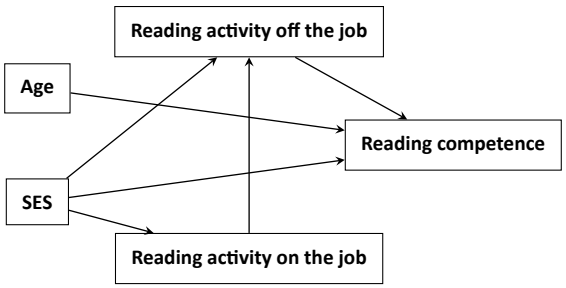

Figure 3: Path diagram of the model.

4.1 Samples for analyses

As in the original study, our sample consisted of adults who were currently employed. We merged the different SUF files for starting cohort 6[3] to create a data set that contained the required data for wave 3 in which the competence tests were administered[4]. This resulted in 5,335 subjects for the plausible values estimation. Although the research question focusses on employed subjects, we chose to use all available test takers in wave 3 to improve measurement precision. After the estimation of plausible values, about 20% of the sample were discarded due

[2] Because the aim of this example is to demonstrate **NEPS*scaling***, we reduced the complexity of the model by omitting education as an independent factor.

[3] For guides on how to merge NEPS data, see the data manuals or merging matrices provided by the Research Data Center of the LIfBi for every starting cohort on https://www.neps-data.de/en-us/datacenter/dataanddocumentation.aspx.

[4] Unfortunately, the amount of reading on and off the job were not administered in wave 5. For other research questions there may be the opportunity of using the additional competence data of wave 5.

to unemployment. Table 4 shows that the complete sample and the subsample of employed subjects have similar socio-demographic properties.

Table 4: Sample characteristics

	Complete sample ($N = 5,335$)		Employed only ($N = 4,303$)	
Variable	*M / Md*	*SD*	*M / Md*	*SD*
Age	48.06	10.87	47.02	9.88
Years of education	14.06	2.41	14.25	2.40
Female[1,2]	0.50	–	0.48	–
Socio-economic status (ISEI-08)	49.30	16.70	49.91	16.68
Migration background[1,2]	0.17	–	0.16	–
Currently employed[1,2]	0.81	–	1.00	–

Notes. The values are the mean values of the descriptive statistics over ten imputed data sets.
[1] Ordered or dichotomous data.
[2] No equals 0, yes equals 1.

4.2 Preparation of background variables

The background variables used for the estimation of plausible values are listed in Table 6. They were chosen for different reasons:

1. Reading activity, age, the International Socio-Economic Index of occupational status (ISEI-08), and the International Standard Classification of Education index (ISCED-97) as an index for the educational level of the subjects were chosen because they are a part of the analysis model (see Figure 3 or Bonerad, 2012).

2. Additional competence domains (e.g., mathematical competence) were included because competence domains are usually highly correlated (e.g., OECD, 2012, 2014, 2017). Inclusion of these variables allows for more precise estimates of plausible values.

3. Similarly, gender and migration background have pronounced effects on German reading competence (Marks, 2008; Verwiebe & Riederer, 2013) and, thus, are expected to improve the plausible values estimation.

This reasoning is in line with the fact that not considering variables that are used in the analysis model (i.e., the SEM in Figure 3) would lead to biased effects between reading competence and the respective variables, whereas considering additional correlates of reading competence in the estimation of plausible values increases the precision of these estimations (Lüdtke & Robitzsch, 2017; Meng, 1994).

Table 6: Variables used in the background model

Variable name	Variable label
maa3_sc1	Mathematics competence (WLE)
mpa3ma_sc5	Procedural meta-cognition (mathematics)
rsa3_sc3	Reading speed
mpa3re_sc5	Procedural meta-cognition (reading)
t34001e_g1	Reading activity (in hours): off the job
t34001f_g1	Reading activity (in hours): on the job
tx29000	Age at wave 3
t400500_g1	Migration background of test target
t700001	Female
tx29063	ISEI-08 (Socio-economic status)
tx29060	Employment status
tx28103	ISCED-97 (Educational level)
tx28102	Years of education

The estimation of the plausible values for the example followed the six steps outlined above:

1. The NEPS SUFs were downloaded from http://www.neps-data.de and stored in the dedicated directory on the local computer.

2. The conditioning variables listed in Table 6 were prepared by the code given in Appendix A. For convenience, this was done in R. However, any other statistical software could have been used as well.

3. The package **NEPS*scaling*** was loaded in the statistical program R. Then the background data prepared in the previous step was imported. In this step, several prerequisites for the use of the function `plausible_values()` should be set:
   - Convert categorical or ordered variables to factors
   - Set the path to the directory with the SUF

```
# load required packages
library(NEPScaling)

# set file path for retrieving data
path <- 'data/SUF SC6 v8/'

# load data (when prepared in R)
load(file = 'data/conditioning_data.RData')

# # load data (e.g. prepared in SPSS)
# # - the file path has to be edited
# # - the package haven also offers functions for Stata or SAS files
# bgdata <- haven::read_spss('data/conditioning_data.sav')

# convert categorical variables into factors
bgdata[, c('gender', 'migration', 'tx29060', 'tx28103')] <-
  lapply(bgdata[, c('gender', 'migration', 'tx29060', 'tx28103')], as.factor)
```

The covariate data is now assigned to the object `bgdata`. The object `path` refers to the path to the SUF directory where the competence data is stored.

4. Estimate plausible values for NEPS starting cohort 6 (`SC = 6`) for reading competence (`domain = 'RE'`) assessed in wave three (`wave = 3`) with `plausible_values()`.

```
# choose the settings for plausible values estimation
# we use the default settings

result <- plausible_values(SC = 6, # starting cohort 6 (adults)
    domain = 'RE', # reading comprehension
    wave = 3, # third wave of NEPS data collection (reading was assessed here)
    path = path, # file path to competence scientific use file of SC 6
    bgdata = bgdata # previously specified background data
)

```

In this example, all optional arguments for the function use the default values (see section 3). All results are stored in the object `result` that contains, next to a record of the arguments passed to `plausible_values()`, a list of `data.frames` of the imputed background variables and plausible values (see Table 2).

5. To analyze the SEM in Figure 3 in another statistical program, the plausible values can be exported into formats readable by the respective programs using the **NEPS*scaling*** function `write_pv()`. The following syntax shows how to save the plausible values and all background variables in SPSS, Stata, and Mplus format.

```
# save plausible values for further analysis in R
save(result, file = 'data/plausible_values.RData')

# save plausible values for further analysis in SPSS
write_pv(pv_obj = result, path = path, ext = 'SPSS')

# save plausible values for further analysis in Stata
write_pv(pv_obj = result, path = path, ext = 'Stata')

# save plausible values for further analysis in Mplus
write_pv(pv_obj = result, path = path, ext = 'Mplus')

```

The plausible values for SPSS, Stata, and Mplus are saved as separate files (e.g., `SC6_RE_w3_cross_plausible_values_1.dat` to `SC6_RE_w3_cross_plausible_values_10.dat` for Mplus) containing the covariate data and one plausible value per subject. Furthermore, for Mplus, a contents file (`content_file.dat`) is generated.

6. Plausible values can be analyzed in any software capable of handling multiply imputed data. We demonstrate these analyses using R and Mplus.

a) We continue to analyze the data in R. The function returns complete data sets that can be used for further analyses. Before we start the actual analyses, we discard all variables not needed for the analysis and select our analysis sample (i.e., all subjects employed in wave 3).

```
analysis_vars <- c('ID_t', 'PV', 'age', 'tx29063', 'tx29060',
                   't34001e_g1', 't34001f_g1')

# load plausible values
load(file = 'data/plausible_values.RData')

# generate analysis data sets
datalist <- lapply(result$pv, function(x) { x[, analysis_vars] })

# keep only target persons who are employed
for (i in 1:length(datalist)) {
  datalist[[i]] <- datalist[[i]][datalist[[i]]$tx29060 == 1, ]
  datalist[[i]]$tx29060 <- NULL
}
```

The data is now sufficiently prepared so that it can be processed by R packages that implement multiple imputation such as **semTools** (semTools Contributors, 2016), **mice** (van Buuren & Groothuis-Oudshoorn, 2011) or **miceadds** (Robitzsch, Grund, et al., 2017). In the following, **lavaan** (Rosseel, 2012) is used for the estimation of the structural equation model and **miceadds**::`pool_mi()` is used to combine the results (Robitzsch, Grund, et al., 2017). The model for the SEM replicating Bonerad (2012, see Figure 3) is specified and passed on to the function `sem_wrapper()`. This user-written function simplifies the syntax. Its code is given in Appendix B, lines 76-100.

```
# model specification for structural equation model
mod <- 'PV ~ t34001e_g1 + age + tx29063
        t34001e_g1 ~ t34001f_g1 + tx29063
        t34001f_g1 ~ tx29063'

# compute SEM and extract standardized parameter estimates
# see appendix B lines 76-100 for definition of the sem_wrapper function
params <- sem_wrapper(datalist)

# pool results of separate SEM analyses
res <- miceadds::pool_mi(qhat = params$qhat, se = params$se)
summary(res)
```

b) Additionally, we conducted the analyses in Mplus. The following code example estimates the SEM as it can be seen in Figure 3. The variable names had to be shortened because Mplus allows only 8 character variable names. This was done manually in the exported data files.

```
TITLE: SEM with plausible values
DATA: FILE IS content_file.dat;
      TYPE = IMPUTATION;
VARIABLE: NAMES ARE isei reoffjob reonjob age employ;
! names as used in R: tx29063 t34001e_g1 t34001f_g1 age tx29060;
      USEVARIABLES ARE isei reoffjob reonjob age;
! keep only employed test takers
      USEOBSERVATIONS = employ EQ 1;
MODEL:
      PV ON reoffjob age isei;
      reoffjob ON reonjob isei;
      reonjob ON isei;
```

Figure 4 shows the pooled standardized results for our example that are given by applying the `summary()` function in R. They can now be interpreted like complete-data estimates and are unbiased on a population level.

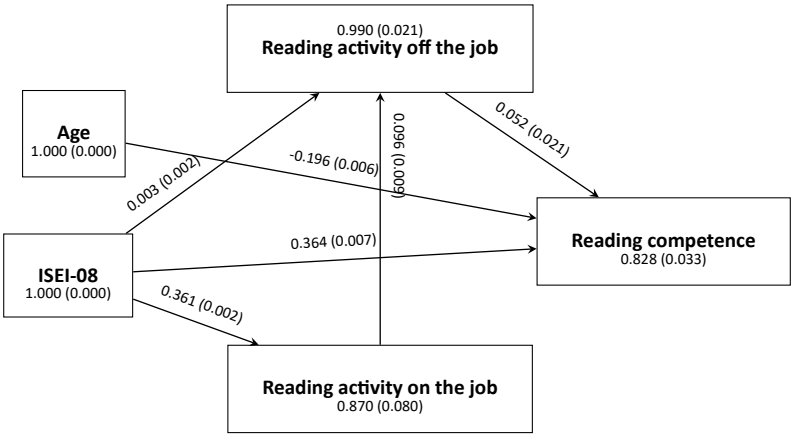

Figure 4: Path diagram of the model. The number inside the nodes signify the pooled variances; the numbers next to the arrows signify pooled standardized regression coefficients. The respective pooled standard error is given in parentheses.

5 Data in the SUFs

For each starting cohort plausible values will be published in the respective scientific use file. The SUF, then, contains 20 plausible values per person, domain and time point. Ten of those are estimated using the cross-sectional model described in this survey paper; the other ten plausible values conform to the longitudinal models. Please note that all estimated models follow the results of NEPS main scalings as closely as possible. This entails fixing the item parameters to the values obtained in the scaling procedure to ensure comparability if identical test forms are administered to multiple starting cohorts (e.g., the assessments of reading and mathematical competences in starting cohorts 4 to 6 in the years 2016 and 2017) and stability of the measurement model when background data is used. Thus, it is recommended to consult the respective technical reports to get an overview of the applied models.

Please note that these plausible values **can only be used in the unlikely case** that the research question at hand contains only (a subset of) the variables listed as the minimal background model in Appendix C[5]. At this point, we urgently advise the researcher to use the R package described in this paper to estimate plausible values themselves **in any other case**. It is essential that the background model used to estimate the plausible values contains the analysis variables.

Furthermore, the R code used to estimate the plausible values is provided as user examples for the respective starting cohort and can be modified for the specific use case. For instance, longitudinal plausible values for reading competence in starting cohort 6 in wave 3 are named "`rea3_pv1u`" to "`rea3_pv10u`" and "`maa3_pv1u`" to "`maa3_pv10u`" for math. Adhering to

[5]Please note that only the model for starting cohort 6 is shown. For further information, see the on-line documentation at https://www.neps-data.de.

NEPS naming conventions, their cross-sectional counterparts are, consequently, named "rea3_pv1" to "rea3_pv10" for reading and "maa3_pv1" to "maa3_pv10" for math.

6 Concluding remarks

This paper introduced the R function plausible_values() in the package **NEPS*scaling*** to generate plausible values for competence tests administered in the NEPS tailored to the specific research question at hand. An important strength of the package is the implemented strategy for handling missing data in the background model. While missing indicators present an easy to use approach that is adopted in many LSAs, recent methodological work provided important advancements in this area and, for example, introduced multiple imputation strategies as alternatives to this simplistic approach (Aßmann et al., 2016; Weirich et al., 2014). Both missing data handling strategies are implemented using CART in **NEPS*scaling***. After introducing the plausible_values() function, we demonstrated the use of the package using the reading competence test in the adult starting cohort. We showed how to analyze SEMs with plausible values in R and Mplus.

Furthermore, we again want to stress that the option min_valid should be manipulated with caution. Following NEPS scaling guidelines (Pohl & Carstensen, 2012), at least three valid responses on a competence test are required to estimate a competence score for a respondent. Still, with the plausible_values() function it is possible to also estimate plausible values for persons that have no test information at all, but only relevant background information or, in the longitudinal case, information on only some of the measurement time points. However, this requires careful model selection and theoretical reasoning regarding the background model.

At the moment, the R package **NEPS*scaling*** is implemented only for starting cohorts 5 and 6; other cohorts will follow shortly. Therefore, it is essential to regularly check for updates of the package as new data releases may result in changes or extensions to the package. The function currently_implemented() returns which starting cohorts, waves and domains are currently supported by **NEPS*scaling***.

Although the paper briefly highlighted the importance of using plausible values for research on population characteristics, we did not present details on how to pool analyses conducted with plausible values using rules of combination for multiple imputation (Little & Rubin, 1987; Mislevy, 1991; Rubin, 1987). However, we want to remind readers that it is of utmost importance to follow these guidelines and not be tempted to use only one of the plausible values in their analyses; respective results are likely severely biased (cf. von Davier et al., 2009).

References

Aßmann, C., Gaasch, C., Pohl, S., & Carstensen, C. H. (2015). Bayesian estimation in IRT models with missing values in background variables. *Psychological Test and Assessment Modeling*, *57*(4), 595–618. https://doi.org/10.1007/978-3-658-11994-2_28 (cit. on p. 8)

Aßmann, C., Gaasch, C., Pohl, S., & Carstensen, C. H. (2016). Estimation of plausible values considering partially missing background information: A data augmented MCMC approach. In H.-P. Blossfeld, J. Skopek, J. Maurice, & M. Bayer (Eds.), *Methodological Issues of Longitudinal Surveys* (pp. 503–521). Springer. (Cit. on p. 22).

Beaton, A. E. (1969). Scaling criterion of questionnaire items. *Socio-Economic Planning Sciences*, *2*(2-4), 355–362. https://doi.org/10.1016/0038-0121(69)90030-5 (cit. on p. 8)

Blossfeld, H.-P., Roßbach, H.-G., von Maurice, J., Schneider, T., Kiesl, S. K., Schönberger, B., Müller-Kuller, A., Rohwer, G., Rässler, S., Prenzel, M. S., et al. (2011). Education as a lifelong process–the German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft: Sonderheft 14*. https://doi.org/10.1007/s11618-011-0179-2 (cit. on p. 5)

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling*, *15*(4), 651–675. https://doi.org/10.1080/10705510802339072 (cit. on p. 7)

Bondarenko, I., & Raghunathan, T. (2016). Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models. *Statistics in Medicine*, *35*(17), 3007–3020. https://doi.org/10.1002/sim.6926 (cit. on p. 7)

Bonerad, E.-M. (2012). *Einflussfaktoren von Leseaktivitäten und Leseverständnis* (PhD). Zürich: Universität. (Cit. on pp. 16, 17, 20).

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1983). *CART: classification and regression trees*. Belmont, CA, Wadsworth. (Cit. on p. 8).

Burgette, L. F., & Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, *172*(9), 1070–1076. https://doi.org/10.1093/aje/kwq260 (cit. on pp. 6, 8, 9)

Doove, L. L., Van Buuren, S., & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, *72*, 92–104. https://doi.org/10.1016/j.csda.2013.10.025 (cit. on pp. 8, 9)

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage Publications Ltd. (Cit. on p. 11).

Fischer, L., Rohm, T., Gnambs, T., & Carstensen, C. H. (2016). *Linking the data of the competence tests*. Bamberg, Germany. (Cit. on pp. 10, 13).

Fuß, D., Gnambs, T., Lockl, K., & Attig, M. (2019). *Competence data in NEPS: Overview of Measures and Variable Naming Conventions (Starting Cohorts 1 to 6)*. Bamberg, Germany. (Cit. on pp. 5, 12).

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, *8*(3), 206–213. https://doi.org/10.1007/s11121-007-0070-9 (cit. on pp. 7, 14)

IBM Corp. (2015). *IBM SPSS Statistics for Windows* [Version 23.0]. Version 23.0. Armonk, NY. (Cit. on p. 6).

Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, *91*(433), 222–230. https://doi.org/10.1080/01621459.1996.10476680 (cit. on p. 8)

Koller, I., Haberkorn, K., & Rohm, T. (2014). *Neps technical report for reading: Scaling results of starting cohort 6 for adults in main study 2012* (tech. rep. No. 48). NEPS Working Paper. Bamberg, Germany. (Cit. on p. 13).

Little, R. J., & Rubin, D. B. (1987). *Statistical analysis with missing data*. John Wiley & Sons. (Cit. on pp. 6, 7, 22).

Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 14–23. https://doi.org/10.1002/widm.8 (cit. on pp. 8, 9)

Lüdtke, O., & Robitzsch, A. (2017). Eine Einführung in die Plausible-Values-Technik für die psychologische Forschung. *Diagnostica*, *63*(3), 193–205. https://doi.org/10.1026/0012-1924/a000175 (cit. on pp. 5, 7, 17)

Marks, G. N. (2008). Accounting for the gender gaps in student performance in reading and mathematics: Evidence from 31 countries. *Oxford Review of Education*, *34*(1), 89–109. https://doi.org/10.1080/03054980701565279 (cit. on p. 17)

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2016). *Methods and procedures in TIMSS 2015*. Boston College, TIMSS & PIRLS International Study Center. http://timssandpirls.bc.edu/publications/timss/2015-methods.html. (Cit. on pp. 5, 8, 13)

Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (2007). *Progress in International Reading Literacy Study (PIRLS): PIRLS 2006 technical report.* ERIC. (Cit. on pp. 8, 13).

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272 (cit. on p. 9)

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 538–558. https://doi.org/10.1214/ss/1177010269 (cit. on pp. 7, 17)

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*(2), 177–196. https://doi.org/10.1007/BF02294457 (cit. on pp. 5, 7, 22)

Muthén, L., & Muthén, B. (1998-2017). *Mplus user's guide* [Eighth Edition]. Eighth Edition. Los Angeles, CA. (Cit. on p. 6).

OECD. (2012). *PISA 2009 technical report*. OECD Publishing. http://dx.doi.org/10.1787/9789264167872-en. (Cit. on p. 17)

OECD. (2014). *PISA 2012 technical report*. OECD Publishing. https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf. (Cit. on p. 17)

OECD. (2015). The abc of gender equality in education: Aptitude, behaviour, confidence. OECD Publishing Paris. (Cit. on p. 6).

OECD. (2017). *PISA 2015 technical report*. OECD Publishing. https://www.oecd.org/pisa/data/2015-technical-report/. (Cit. on pp. 5, 8, 13, 17)

Pohl, S., & Carstensen, C. H. (2012). *NEPS technical report – Scaling the data of the competence tests*. Otto-Friedrich-Universität, National Educational Panel Study. Bamberg. (Cit. on pp. 5, 6, 9–11, 13, 16, 22).

Rasch, G. (1960). Probabilistic models for some intelligence and achievement tests. *Copenhagen: Danish Institute for Educational Research* (cit. on pp. 6, 10).

Robitzsch, A., Grund, S., & Henke, T. (2017). *Miceadds: Some additional multiple imputation functions, especially for 'mice'* [R package version 2.4-12]. R package version 2.4-12. https://CRAN.R-project.org/package=miceadds. (Cit. on p. 20)

Robitzsch, A., Kiefer, T., & Wu, M. (2017). *TAM: Test analysis modules* [R package version 2.6-2]. R package version 2.6-2. https://CRAN.R-project.org/package=TAM. (Cit. on pp. 10, 11)

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. http://www.jstatsoft.org/v48/i02/ (cit. on p. 20)

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY, Wiley. https://doi.org/10.1002/9780470316696. (Cit. on pp. 6, 7, 22)

Rutkowski, L. (2011). The impact of missing background data on subpopulation estimation. *Journal of Educational Measurement*, *48*(3), 293–312. https://doi.org/10.1111/j.1745-3984.2011.00144.x (cit. on p. 8)

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147. https://doi.org/10.1037/1082-989X.7.2.147 (cit. on p. 8)

semTools Contributors. (2016). *semTools: Useful tools for structural equation modeling* [R package version 0.4-14]. R package version 0.4-14. https://CRAN.R-project.org/package=semTools. (Cit. on p. 20)

StataCorp. (2015). *Stata Statistical Software* [Release 14]. Release 14. College Station, TX. (Cit. on p. 6).

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03 (cit. on p. 20)

Verwiebe, R., & Riederer, B. (2013). Die Lesekompetenzen von Jugendlichen mit Migrationshintergrund in westlichen Gesellschaften/The Reading Literacy of Immigrant Youth in Western Societies. *Zeitschrift für Soziologie*, *42*(3), 201–221. https://doi.org/10.1515/zfsoz-2013-0303 (cit. on p. 17)

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful. *IERI Monograph Series*, *2*, 9–36 (cit. on pp. 5, 7, 22).

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. https://doi.org/10.1007/BF02294627 (cit. on p. 5)

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen, C. H. (2011). Development of competencies across the life span. *Zeitschrift für Erziehungswissenschaft*, *14*(2), 67–86. https://doi.org/10.1007/s11618-011-0182-7 (cit. on p. 5)

Weirich, S., Haag, N., Hecht, M., Böhme, K., Siegle, T., & Lüdtke, O. (2014). Nested multiple imputation in large-scale assessments. *Large-scale Assessments in Education*, *2*(1), 9. https://doi.org/10.1186/s40536-014-0009-0 (cit. on pp. 10, 22)

Wickham, H., Francois, R., Henry, L., & Müller, K. (2017). *Dplyr: A grammar of data manipulation* [R package version 0.7.4]. R package version 0.7.4. https://CRAN.R-project.org/package=dplyr. (Cit. on p. 11)

Wickham, H., & Miller, E. (2016). *Haven: Import and export 'SPSS', 'Stata' and 'SAS' files* [R package version 1.0.0]. R package version 1.0.0. https://CRAN.R-project.org/package=haven. (Cit. on p. 11)

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*(2-3), 114–128. https://doi.org/10.1016/j.stueduc.2005.05.005 (cit. on pp. 5, 7)

Yamamoto, K., Khorramdel, L., & Von Davier, M. (2013). Scaling PIAAC cognitive data. *Technical report of the survey of adult skills (PIAAC)*, 408–440 (cit. on p. 8).

Zinn, S., & Gnambs, T. (2018). Modeling competence development in the presence of selection bias. *Behavior Research Methods*, 1–16. https://doi.org/10.3758/s13428-018-1021-z (cit. on pp. 5, 7)

A R code used for the preparation of the example covariate data

```
rm(list = ls(all.names = TRUE)); gc()

# prepare conditioning data for working paper example
# - from Bonderad's dissertation (2012)
# - variables:
#   - reading on the job (t34001f_g1)
#   - reading off the job (t34001e_g1)
#   - educational level (ISCED-97) (tx28103)
#   - years of education (tx28102)
#   - age (tx29000) -- from Bascis v3
#   - ISEI-08 (tx29063) -- from Bascis v3
#   - gender (t700001)
#   - employment status (tx29060) -- from Bascis v3
#   - migration background (t400500_g1)
#   - maths wle (maa3_sc1)
#   - procedural meta-cognition (maths) (mpa3ma_sc5)
#   - procedural meta-cognition (reading) (mpa3re_sc5)
#   - reading spead (rsa3_sc3)
# - data sets:
#   - pTarget
#   - Basics
#   - xTargetCompetencies


# load packages
library(haven);library(dplyr)

# set file path for retrieving data
path <- 'data/SUF SC6 v8/'

# read in respective data files
pTarget <- read_spss(paste0(path,'SC6_pTarget_D_8-0-0.sav'))
xTargetCompetencies <- read_spss(paste0(path,'SC6_xTargetCompetencies_D_8-0-0.sav'))
genstat <- read_spss(paste0(path,'SC6_Basics_D_8-0-0.sav'))
# basics from wave 3
Basics <- read_spss('data/SUF SC6 v3/SC6_Basics_D_3-0-1.sav')
Basics <- Basics[, c('ID_t', 't700001', 'tx29000', 'tx28103', 'tx28102', 'tx29060',
    'tx29063')]
Basics <- left_join(Basics, genstat[, c('ID_t', 't400500_g1')], by = 'ID_t')
rm(genstat)

# extract competence measures
competencies <- xTargetCompetencies[xTargetCompetencies$wave_w3 == 1, c('ID_t', '
    maa3_sc1', 'mpa3ma_sc5', 'mpa3re_sc5', 'rsa3_sc3')]
rm(xTargetCompetencies)

# extract reading activity
pTarget <- pTarget[, c('ID_t', 'wave', 'splink', 't34001e_g1', 't34001f_g1')]


# combine information to one data set
# - merge Basics to pTarget
pTarget <- pTarget[order(pTarget$ID_t) & pTarget$wave == 3, c('ID_t', 'wave', '
    t34001e_g1', 't34001f_g1')]
BpT <- full_join(Basics, pTarget, by = 'ID_t')
rm(Basics, pTarget)

# - clean data
bgdata <- BpT[order(BpT$ID_t)
, c('ID_t', 'tx29000', 't700001', 't400500_g1'
, 'tx29060', 'tx29063', 't34001e_g1', 't34001f_g1', 'tx28103', 'tx28102')]
rm(BpT)
bgdata <- full_join(bgdata, competencies, by = 'ID_t')
```

```
rm(competencies)
bgdata$gender <- bgdata$t700001 - 1
bgdata$migration <- ifelse(bgdata$t400500_g1 == 0, 0, 1)
bgdata$age <- floor(bgdata$tx29000)

psych::describe(bgdata)

bgdata$t700001 <- bgdata$t400500_g1 <- bgdata$tx29000 <- NULL


# save data
save(bgdata, file = 'data/conditioning_data.RData')

```

B R code of the example en bloc

```


# estimate plausible values for SC6 reading comprehension (wave 3)
# load required packages
library(NEPScaling)

# set file path for retrieving data
path <- 'data/SUF SC6 v8/'

# load data (when prepared in R)
load(file = 'data/conditioning_data.RData')

# # load data (e.g. prepared in SPSS)
# # - the file path has to be edited
# # - the package haven also offers functions for Stata or SAS files
# bgdata <- haven::read_spss('data/conditioning_data.sav')

# convert categorical variables into factors
bgdata[, c('gender', 'migration', 'tx29060', 'tx28103')] <-
    lapply(bgdata[, c('gender', 'migration', 'tx29060', 'tx28103')], as.factor)

# choose the settings for plausible values estimation
# we use the default settings
result <- plausible_values(SC = 6, domain = 'RE', wave = 3,
    path = path, bgdata = bgdata,
    control = list(ML = list(itermcmc = 100, burnin = 50)))

# save plausible values for further analysis in R
save(result, file = 'data/plausible_values.RData')

# save plausible values for further analysis in SPSS
write_pv(pv_obj = result, path = path, ext = 'SPSS')

# save plausible values for further analysis in Stata
write_pv(pv_obj = result, path = path, ext = 'Stata')

# save plausible values for further analysis in Mplus
write_pv(pv_obj = result, path = path, ext = 'Mplus')

# replicate part of Bonderad's (2012) dissertation project
# variables of the conditioning model used in the analysis:
# - reading on the job (t34001f_g1)
# - reading off the job (t34001e_g1)
# - age (age)
# - ISEI-08 (tx29063)
analysis_vars <- c('ID_t', 'PV', 'age', 'tx29063', 'tx29060', 't34001e_g1',
                   't34001f_g1')

# load plausible values
```

```
load(file = 'data/plausible_values.RData')

# generate analysis data sets
datalist <- lapply(result$pv, function(x) { x[, analysis_vars] })

# keep only target persons who are employed
for (i in 1:length(datalist)) {
datalist[[i]] <- datalist[[i]][datalist[[i]]$tx29060 == 1, ]
datalist[[i]]$tx29060 <- NULL
}
rm(i)

# model specification for structural equation model
mod <- 'PV ~ t34001e_g1 + age + tx29063
        t34001e_g1 ~ t34001f_g1 + tx29063
        t34001f_g1 ~ tx29063'

# compute SEM and extract standardized parameter estimates
# see appendix B lines 76-100 for definition of the sem_wrapper function
params <- sem_wrapper(datalist)

# pool results of separate SEM analyses
res <- miceadds::pool_mi(qhat = params$qhat, se = params$se)
summary(res)

rm(mod, fit, qhat, se, analysis_vars)

#' sem_wrapper function: has to be run before SEM estimation;
#' package lavaan has to be installed
#' @param datalist list of multiply imputed data sets
#' @return list of test statistic and respective standard error
sem_wrapper <- function(datalist) {
  fit <- lapply(datalist, FUN = function(data){
    res <- lavaan::sem(mod,data = data)
    return(res)
  })
  qhat <- lapply( fit , FUN = function(ll){
    h1 <- lavaan::parameterEstimates(ll, standardized = TRUE)
    parnames <- paste0( h1$lhs , h1$op , h1$rhs )
    v1 <- h1$std.all
    names(v1) <- parnames
    return(v1)
  } )
  se <- lapply( fit , FUN = function(ll){
    h1 <- lavaan::parameterEstimates(ll, standardized = TRUE)
    parnames <- paste0( h1$lhs , h1$op , h1$rhs )
    v1 <- h1$se
    names(v1) <- parnames
    return(v1)
  } )
  return(list(qhat=qhat, se=se))
}
```

C Minimal background model for starting cohort 6

*Table 8: Minimal background model for **starting cohort 6**.*

Variable name	Description	Data set
t700001	gender	Basics
t405000_g2	state of birth	Basics
t70000y[1]	year of birth	Basics
tx80101	federal state	Methods
tx80102	BIK category (size of town)	Methods
tx28101	CASMIN	Education
t34005a[2]	number of books at home	pTarget
t400500_g1[3]	generation status	pTarget
ts23901	employment status	spEmp
tx80107	subsample (ALWA or NEPS)	Methods
rea3_sc1[4]	WLE for reading competence wave 3	xTargetCompetencies
rea5_sc1[4]	WLE for reading competence wave 5	xTargetCompetencies
rea9_sc1[4]	WLE for reading competence wave 9	xTargetCompetencies
maa3_sc1[4]	WLE for mathematical competence wave 3	xTargetCompetencies
maa9_sc1[4]	WLE for mathematical competence wave 9	xTargetCompetencies
sca5_sc1[4]	WLE for ICT literacy wave 5	xTargetCompetencies
ica5_sc1[4]	WLE for science competence wave 5	xTargetCompetencies
mpa3re_sc5	procedural meta-cognition for reading competence wave 3 (difference measure)	xTargetCompetencies
mpa5re_sc5	procedural meta-cognition for reading competence wave 5 (difference measure)	xTargetCompetencies
mpa9re_sc5	procedural meta-cognition for reading competence wave 9 (difference measure)	xTargetCompetencies
mpa3ma_sc5	procedural meta-cognition for mathematical competence wave 3 (difference measure)	xTargetCompetencies
mpa9ma_sc5	procedural meta-cognition for mathematical competence wave 9 (difference measure)	xTargetCompetencies
mpa5ic_sc5	procedural meta-cognition for ICT literacy wave 5 (difference measure)	xTargetCompetencies
mpa5sc_sc5	procedural meta-cognition for science competence wave 5 (difference measure)	xTargetCompetencies
nr	number of not-reached items	generated by NEPScaling

Notes. All variables used in the minimal background model are constant over time or change rarely. They have been shown to correlate with the competencies in question and some of them are also used in the NEPS weighting or scaling procedures. For competencies other than reading, the WLE for reading competence should be added and the respective WLE removed. In the longitudinal case, proxies for all competencies but the one to be measured are added to the background model as well. To avoid inconsistencies over time, only WLEs for the competence at the *first* measurement point are used.

[1] In the cross-sectional case, the age at the time of measurement is computed from this variable and used instead.

[2] The number of books at home is dichotomized in NEPS scaling into less than (corresponding to the values 1 to 3) and at least 100 books.

[3] A generation status of 0 or more than 2.25 generations back (values 0 or at least 4) is considered as not having a migration background in NEPS scaling.

[4] In the longitudinal case, uncorrected WLEs are used as competence proxies. Their variable name is extended with the letter "u".