Scharl and Zink
Large-scale Assessments in Education (2022) 10:28
https://doi.org/10.1186/s40536-022-00145-5

Large-scale Assessments in Education

RESEARCH Open Access

NEPSscaling: plausible value estimation for competence tests administered in the German National Educational Panel Study

Anna Scharl and Eva Zink*

*Correspondence: eva.zink@lifbi.de

Leibniz Institute for Educational Trajectories, Wilhelmsplatz 3, 96047 Bamberg, Germany

Abstract

Educational large-scale assessments (LSAs) often provide plausible values for the administered competence tests to facilitate the estimation of population effects. This requires the specification of a background model that is appropriate for the specific research question. Because the *German National Educational Panel Study* (NEPS) is an ongoing longitudinal LSA, the range of potential research questions and, thus, the number of potential background variables for the plausible value estimation grow with each new assessment wave. To facilitate the estimation of plausible values for data users of the NEPS, the R package NEPS*scaling* allows their estimation following the scaling standards in the NEPS without requiring in-depth psychometric expertise in item response theory. The package requires the user to prepare the data for the background model only. Then, the appropriate item response model including the linking approach adopted for the NEPS is selected automatically, while a nested multiple imputation scheme based on the chained equation approach handles missing values in the background data. For novice users, a graphical interface is provided that only requires minimal knowledge of the R language. Thus, NEPS*scaling* can be used to estimate cross-sectional and longitudinally linked plausible values for all major competence assessments in the NEPS.

Keywords: National Educational Panel Study, Plausible values, Competence, NEPS*scaling*, Large-scale assessment

Introduction

For decades, educational large-scale assessments (LSAs) have provided insights into educational systems around the globe (e.g., PISA, NAEP, TIMSS, and PIRLS). Usually, these LSAs are cross-sectional and study specific age cohorts (e.g., 15 year olds in the case of PISA; Weis & Reiss, 2019). Although repeated assessment cycles allow longitudinal comparisons on the country level, LSAs providing access to within-person change trajectories are rare. In contrast, the *National Educational Panel Study* (NEPS) is a longitudinal, multi-cohort study representative for the German population. The NEPS follows newborns to pensioners through repeated assessments across their life courses (Blossfeld & von Maurice, 2011). Currently, the NEPS includes four child cohorts of

newborns (starting cohort 1), kindergartners (starting cohort 2), fifth graders (starting cohort 3), and ninth graders (starting cohort 4) as well as two grown-up cohorts of university students (starting cohort 5) and adults between 30 and 70 years old (starting cohort 6). A major focus of the NEPS is the coherent measurement of domain-specific competencies such as reading, math, or sciences across all cohorts to study prerequisites and outcomes of education in Germany. In LSAs, competencies are typically analyzed as plausible values (PVs) (Wu, 2005). PVs allow for unbiased population parameter estimation at the population level as they take the uncertainty of the estimation of the latent competences into account by providing multiple estimates representing the likely distribution of the competences (Lechner et al., 2021; Lüdtke & Robitzsch, 2017). It is important to note that PVs are a special case of multiple imputation, and therefore, the assumptions behind the multiple imputation approach need to be met (Rubin, 1987). Precise PV estimation requires the specification of a background model that is appropriate for the research question at hand. Hereby, all variables (and their interactions) used in later analyses need to be included (Bondarenko & Raghunathan, 2016; Meng, 1994). Because data providers of LSAs cannot anticipate how users will analyze their data, typically all available information collected in a LSA is incorporated into the PV estimation to achieve said unbiasedness (e.g. OECD, 2017). A challenge of longitudinal LSAs such as the NEPS is their growing data base. Each new assessment wave needs to be incorporated into earlier PV estimation to accommodate PVs as independent variables in all possible statistical models (cf. congenial models; Meng, 1994). Paired with the need for completely observed background data, the estimation of ready-to-use PVs in scientific use files (SUFs) quickly becomes impractical. Therefore, we introduce the R package NEPS*scaling* that offers versatile functionalities to estimate PVs for cross-sectional and longitudinally linked competence assessments in each cohort of the NEPS. Although, plausible values can also be generated with other available R packages such as TAM (Robitzsch et al., 2021), mirt (Chalmers, 2012) or brms (Bürkner, 2017) as well as other standalone software such as Mplus ((Muthén & Muthén, 1998), what sets NEPS*scaling* apart from these software packages is its scope and simplicity. It is designed to specifically suit the needs of the NEPS and its data users. Therefore, the package is also aimed at researchers with little expertise in psychometric modeling and novice users of R. It only requires the preparation of custom background data that fits the research question, which can be done with any statistical software as long as the data is exported in an R readable way (e.g., CSV, SPSS, Stata, SAS formats). Then, the package automatically handles missing values in the background data using classification and regression trees (CART) in a nested imputation scheme (Burgette & Reiter, 2010; Weirich et al., 2014). It estimates the appropriate item response models following the scaling standards in the NEPS (Pohl & Carstensen, 2013) to generate PVs suiting the intended analyses. Finally, the generated data can be exported in different formats for various statistical software such as SPSS, Stata, or Mplus. Furthermore, a graphical user interface is provided for novice R users.

In the following, we will briefly outline the statistical background of PVs and then describe the basic functionality of NEPS*scaling*. The use of the package is demonstrated in two examples that show how to estimate PVs using either R syntax or the graphical user interface.

Background

Plausible values

Following the NEPS*scaling* standards (Pohl & Carstensen, 2013), most competence tests are scaled using the partial credit model (PCM; Masters, 1982) for polytomous items which models the probability of observing response Y_{ij} for person i on item j as

$$P(Y_{ij} = y|\theta_i, \delta_j) = \frac{\exp\{y \cdot \theta_i - \sum_{k=0}^{y} \delta_{jk}\}}{\sum_{h=0}^{K_j} \exp\{h \cdot \theta_i - \sum_{k=0}^{h} \delta_{jk}\}} \quad \text{with} \quad \delta_{j0} = 0 \tag{1}$$

where θ_i denotes the latent ability of person i and δ_{jk} the threshold for endorsing category $k = \{0, \ldots, K_j\}$ of item j. It simplifies to the Rasch model (Rasch, 1960) in case of binary items. For rotated test designs that administered a given test at different positions for the sample, a multi-facet model (Linacre, 1989) based on the PCM or Rasch model is used to correct for the test rotation.[1] Moreover, longitudinal assessments are linked across measurement waves using mean/mean linking (Fischer et al., 2019) which shifts the latent scale to be anchored at the first measurement's mean location.

The PVs technique is an extension of IRT models via a latent regression of the person parameters on background variables (Lüdtke & Robitzsch, 2017). This allows to approximate the population level latent distribution of person abilities more accurately. The latent regression of θ_i can be seen as prior information on the person parameters and leads to the formulation of the posterior ability distribution of person i as

$$p(\theta_i|\boldsymbol{y_i}, \boldsymbol{x_i}) \propto p(\boldsymbol{y_i}|\theta_i)p(\theta_i|\boldsymbol{x_i}) \tag{2}$$

where $p(\boldsymbol{y_i}|\theta_i)$ denotes the likelihood of the data, given by the IRT model, and $p(\theta_i|\boldsymbol{x_i})$ denotes the prior distribution of the latent ability, given by the latent regression on a set of background variables $\boldsymbol{x_i}$ for person i

$$\theta_i = \beta_0 + \boldsymbol{x_i}\boldsymbol{\beta_L} + \boldsymbol{\varepsilon_i} \tag{3}$$

with $\beta_L = (\beta_1, \ldots, \beta_L)^T$ denoting the regression weights for L covariates, the intercept β_0, and ε_i representing the normally distributed residual. The latent regression should contain all relevant variables and variable configurations such as interactions or non-linear terms that are part of the planned analyses (Bondarenko & Raghunathan, 1994; Meng, 2016). Similarly, it may be sensible to add further variables to improve the imputation of the background data (Collins et al., 2001), especially if it is not predictable which variable relations are of interest in later analyses (Lüdtke & Robitzsch, 2017).

This also highlights that PVs are a special case of multiple imputation of completely missing variables. Therefore, analyses with PVs have to be conducted separately for each single PV and then combined following Rubin's rules (Lechner et al., 2021; Rubin, 1987). Reaserchers need to check whether the assumption of normally distributed parameter estimates holds, before pooling the datasets using Rubin's rules.

[1] Multiple competence tests administered in the same wave are typically presented in different sequence to respondents in order to balance potential fatigue effects across the different tests. Because this might distort between-respondent comparisons, cross-sectional analyses should correct for the adopted rotation design. In contrast, longitudinal analyses typically focus on within-person comparisons. Because for a given respondent the test position remains constant across different measurement waves, usually no corrections for the test rotation are necessary. Therefore, the multi-facet model correcting for the test rotation is only applied to the estimation of cross-sectional plausible values.

Classification and regression trees

The missing data strategy in LSAs for plausible values estimation typically encompasses re-defining missing values due to item nonresponse as an additional dummy variable during the recoding of the background data. This cannot be seen as effectively handling missing data (Lüdtke et al., 2002; Schafer & Graham, 2017) which is why we adopted nested multiple imputation (Weirich et al., 2014). This strategy first repeatedly imputes the background data and then estimates the desired number of PVs for each imputed data set. Additionally, it can consider dependencies between the ability and the background variables if an ability indicator like the weighted likelihood estimate (Warm, 1989) is used in the imputation model. Please note that this strategy does not apply to unit nonresponse, that is, if a case is completely unobserved in an assessment wave. These cases are handled by listwise deletion.

The background model is imputed using a CART algorithm within the multiple imputation via chained equations (MICE) framework (Burgette & Reiter, 2010; Doove et al., 2014). The algorithm predicts a missing value on one variable by a set of predictor variables. Starting with all non-missing values of the outcome variable, the algorithm recursively splits the nodes into binary partitions until a purity criterion is met, that is, the values left in the leaf nodes of the tree are homogeneous enough (Burgette & Reiter, 2010). If the outcome variable is metric, a regression tree is constructed. It differs from a classification tree for categorical outcomes in its purity criterion and the way, a final value is chosen from the respective leaf nodes. A notable advantage of the non-parametric CART as compared to other parametric imputation approaches, like predictive mean matching (Little, 1988) or fully conditional specification (Raghunathan et al., 2003; van Buuren, 2006; van Buuren et al.,2012), is that the splitting of child nodes automatically implies non-linear relationships as well as interactions in the data without having to explicitly model them (Burgette & Reiter, 2010). Further, also monotone transformations of the independent variables do not affect the trees produced by CART (Breiman et al., 2017).

About NEPS*scaling*

NEPS*scaling* is an R package containing functions to facilitate the estimation of PVs for competence domains measured in the NEPS while handling missing values in the background model. Other functions allow the inspection of the specific CARTs used for imputation, accessing parts of the resulting NEPS*scaling* R object, and exporting PVs for different statistical software. The user can also call information about the implemented competences tests and assessment waves, as well as differences between the package and the pre-calculated competence measures published in the NEPS scientific-use files (SUFs). The SUFs include the prepared survey and test data in a factually anonymized form. To download the SUFs for research purposes, only a data use agreement needs to be signed.

Availability

The package NEPS*scaling* is not available from CRAN, but is provided by the NEPS research data center at https://www.neps-data.de/Data-Center/Overview-and-Assis

tance/Plausible-Values. The package is free to download and previous package versions and example code that shows how to use the package to estimate PVs in different cohorts are also available. To install the package, the following command can be used

```
install.packages("NEPSscaling",
    repos=c("http://nocrypt.neps-data.de/r",
    "https://cran.r-project.org"))
```

Basic functions

In the following the most important functions are described in the order in which they would occur during a typical use of the package.

The functions *currently_implemented()* and *deviations_of_package_from_suf()* require no arguments and give an overview of the current state of NEPS*scaling*. The former shows which competence tests are available for which starting cohort by domain and wave, while the latter reports known deviations in comparison to the point estimates of the latent competences (WLEs) provided in the NEPS SUFs.

The main function for generating PVs is *plausible_values()* which loads the raw data from the scientific use files, imputes missing values in the background data, creates the appropriate scaling model for the chosen competence test, and estimates either cross-sectional or longitudinally linked PVs.

The function expects several arguments specifying the PV estimation; most of them are optional:

- *SC* (required): The starting cohort is given by specifying its integer equivalent (e.g., the adult cohort is listed as starting cohort 6).
- *domain* (required): The chosen competence domain is indicated by the two or three letter abbreviations summarized in Fuß et al. (2021). Because not all competence domains have been assessed in each cohort, users have to specify the correct combination of *SC* and domain as indicated by *currently_implemented()*.
- *wave* (required). The assessment wave is given by an integer value as summarized in Fuß et al. (2021). For example, the tests of the starting cohort 6 (adults) took place in the waves 3, 5, and 9.
- *path* (required): Because the function automatically loads the relevant data from the scientific use files, the path to the data on the hard drive needs to be specified as a string (e.g., "C:/Users/name/NEPS_data/" on a Windows machine).
- *bgdata* (optional): The background data needs to be provided as a *data.frame* containing the person identifier *ID_t*. If no background data is provided, PVs without a background model are estimated. Note that the package automatically includes the number of not-reached missing values as a proxy for processing times and, if the assessment took place in the school context, the mean competence per school as a proxy for the multi-level sampling design; this default setting can be changed using the arguments *include_nr* and *approximate_school_context* explained below.

- *npv* (optional): The number of randomly drawn PVs can be explicitly set, but defaults to a value of 10. Importantly, only *npv* PVs are returned even if more sets are estimated.
- *nmi* (optional): The number of randomly drawn imputed background data sets can be explicitly set, but defaults to a value of 10.
- *min_valid* (optional): PVs are only estimated for respondents that provided a minimum number of valid (i.e., non-missing) responses (default: 3).
- *longitudinal* (optional): The logical argument indicates whether cross-sectional PVs for the specified wave or longitudinally linked PVs of the specified cohort should be estimated. In the longitudinal case, all available waves of the specified cohort are included in the estimation.
- *rotation* (optional): The logical argument indicates whether the test rotation design should be considered in the cross-sectional case, thus, estimating a multi-faceted model.
- *include_nr* (optional): The logical argument specifies whether the number of not-reached items (i.e., missing values) should be included in the background model as a proxy for test taking effort.
- *adjust_school_context* (optional): The logical argument controls whether the average point estimate (WLE) of each school of the competence should be calculated and included in the background model to approximate the nested sampling scheme in school assessments.
- *exclude* (optional): Some variables included in the supplied background data can be excluded from the estimation model of the PVs and only be used for imputing missing values. In the cross-sectional case, the argument is a character vector (e.g., *c("var1", "var3"))*, while it must be a named list (e.g., *exclude = list(w1 = "var1", w3 = c("var1", "var3"))*) in the longitudinal case specifying the excluded variables for each wave.
- *seed* (optional): For reproducibility, the specific seed can be set for the random number generators.
- *control* (options): The list can contain logicals informing whether point estimates in the form of WLEs and expected a posteriori estimates (EAPs) should be returned. Additional arguments are passed on to the estimation algorithm in TAM's *tam.mml()* and *tam.pv()* functions. List of additional options: If *EAP = TRUE*, the EAPs will be returned as well; for *WLE = TRUE*, WLEs are returned. Furthermore, additional control options for are collected in the list 'ML'. 'minbucket' defines the minimum number of observations in any terminal CART node (defaults to 5), 'cp' determines the minimum decrease of overall lack of fit by each CART split (defaults to 0.0001).

.

After the estimation of PVs, the functions *print(x)* and *summary(object)* give a quick overview of the specified model and estimated model parameters. The only required argument is the R object resulting from using *plausible_values()*. To facilitate the exploration of the resulting R object, the package also contains a number of extraction functions such as *get_domain(pv_obj), get_info_criteria(pv_obj), get_pv_list(pv_obj), or get_pv_index(pv_obj, index)*. Moreover, the CART imputation can be visualized with

display_imputation_tree(pv_obj, imputation, variable) that generates a plot displaying the specific tree constructed to impute a single variable. If the graphical representation becomes too complex, a character representation of the tree can be inspected using *get_imputation_tree(pv_obj, imputation, variable)*.

The package also provides means to easily export the estimated PVs together with their imputed background data in case analyses with the PVs are to be conducted using different software. The *write_pv()* function takes the arguments *pv_obj*, that is, the resulting R object, *path* where the data is to be stored and *ext*, a string indicating the storage format (i.e., SPSS, Stata, or Mplus).

Typical workflow

Estimating PVs for competence tests in the NEPS typically follows several consecutive steps that depend on two data sources. First, the SUFs including the raw competence test data need to be obtained from https://neps-data.de. Data access requires a free, non-commercial data use agreement with the NEPS research data center.[2] This data is necessary to estimate the IRT part of the plausible values model and needs to be stored in a way that it is accessible to the current R session. Ideally, all raw data files can be found in the same folder. Second, the carefully selected background variables for the PV estimation need to be prepared by the user to ensure congeniality with the intended analyses. This data preparation can be done using any statistical software and the user is advised to check the data for plausibility before starting any further analyses. The only requirement for using NEPS*scaling* is that the resulting background data is stored in tabular format either in SPSS's sav format, Stata's dav, or R's rds format (when using the graphical user interface) or that it is imported into the current R session as a *data.frame* (when invoking the estimation using the R script). Special attention needs to be paid to missing data in the background data as they must be coded as R's *NA* values and categorical variables have to be converted to R's *factors*. Further, the selected background variables should either be assessed at the same time point for which the PVs will be estimated or include time constant information to avoid inconsistencies. After the preparation of the background data, PVs can be estimated via an R script and the functions outlined above or via the graphical user interface provided by the NEPS*shiny* app.

Last, NEPS*scaling* versions always depend on different versions of the SUFs because the competence variables in the SUFs are addressed by the package's functions. Therefore, if variable names are changed in the SUFs, they are changed accordingly in the newest version of the package. As a consequence, the names of newer SUF versions and older package versions and vice versa are no longer compatible. Thus, it is recommended to always use the latest versions of both SUF and package to ensure a match. Further, it is advised to state package versions explicitly to ensure reproducibility.

Surrounding NEPS*scaling*

The estimation of PVs can and should be conducted with several things in mind: First, the imputation model may unduly impact the results and thus, conducting sensitivity

[2] https://www.neps-data.de/Data-Center/Data-Access/Data-Use-Agreements.

analyses for different imputation models or evaluating the efficiency gains obtained through using PVs is advisable. Second, there are particularities in working with plausible values. For example, it is necessary to conduct any further analyses (e.g., regression analysis) which use PVs separately for each set of PVs. The results of these analyses then need to be pooled using Rubin's rules or any other appropriate pooling procedure (Raghunathan et al., 2003). For further information how to correctly work with PVs see von Davier et al. (2009) and for further information on the estimation of PVs as well as an example for pooling the results of Scharl et al. (2020).

Applications

In the following, two example applications are presented that use simulated data sets included in the package. The data was modeled to closely resemble the adult starting cohort (SC 6) and the 5th grader starting cohort (SC 3). The first example will be presented using a classic R script, whereas the second example uses the NEPS*scaling* Shiny app. The aim of the presented applications is to demonstrate basic analyses. Real and more complex examples can be found in further user examples given at the download site of NEPS*scaling* as well as simulated examples for background data.

The input data in both example applications is dictated by the NEPS SUF format. The SUFs are available as SPSS or Stata tables. NEPS*scaling* uses the competence data as it was downloaded. The background data, on the other hand, needs to be prepared by the user as described in chapter 3.2 and should contain the set of analysis variables as well as optional further variables that would improve the imputation of missing background values or the estimation of PVs. NEPS*scaling* internally selects only those subjects in the background data set who have contributed at least the minimum number of valid responses in the competence test of interest.

Application 1: Cross-sectional reading competence in the adult starting cohort

Estimating plausible values using an R script is straightforward. After preparing the background data in any statistical program and storing the background data in one of the supported file formats, there are three steps until PVs are ready for further processing. The first step consists of installing NEPS*scaling*, setting the workling directory for importing the prepared background data into R and loading NEPS*scaling*. Then, PVs can be estimated. It is important to specify the correct path to the competence data. Here, the competence data is stored in a folder called SC6. However, the name of the folder can be chosen freely as long as the path is specified correctly.

```
library(NEPSscaling)
setwd()
bgdata <- readRDS("bgdata.rds")
pv_obj <- plausible_values(SC = 6, domain = "RE", wave = 3,
    path = "./SC6/", bgdata = bgdata)
summary(pv_obj)
```

Below, the abbreviated summary of the estimated model is given. It contains the basic parameters of the estimated model, mean, variance and reliability estimates of the PVs, the fixed item difficulties, and the estimated latent regression weights. The latter cannot

be used to answer the intended research questions, but are giving insights into the influence of the chosen background variables on the estimation of the PVs. They are not meant to be used in further analyses.

```
## Plausible Values Estimation with NEPSscaling
##
## Starting Cohort:  6
## Domain:  RE
## Wave(s):  3
## Test takers per wave:  3000
## Number of estimated plausible values:  10
## Number of sampled imputations / completed data:  1
##
## EAP reliability:  0.804
##
## Variables in background model:  age2, gender2, nbooks2,
   migration2
##
## Starting time:  2021-10-04 14:20:31
## Time for estimation:  23.6 secs
## Total computation time:  23.9 secs
##
## Mean of Plausible Values:
##     PV
## -1.426
##
## Variance of Plausible Values:
## [1] 0.681
##
## Item parameters:
##                      xsi se.xsi
## rea30110_c        -3.605  0.000
## rea3012s_c        -1.605  0.000
[...]
## rea30550_c         0.107  0.000
## position1         -0.003  0.005
## rea3012s_c:step1 -0.105  0.053
## rea3015s_c:step1  0.041  0.049
[...]
## rea3052s_c:step5  0.562  0.075
## rea3054s_c:step5 -0.816  0.073
##
## Regression Coefficients:
##     Variable imp1_coeff imp1_coeff_std imp1_se
## 1  Intercept      0.000             NA   0.000
## 2       age2      0.462          0.279   0.025
## 3    gender2      0.000          0.000   0.024
## 4    nbooks2      0.498          0.300   0.025
## 5 migration2     -0.063         -0.027   0.044
```

In a final step, the plausible values and the imputed background data can be exported for further analysis (here: SPSS file format).

```
write_pv(pv_obj, path = "/SC6", ext = "SPSS")
```

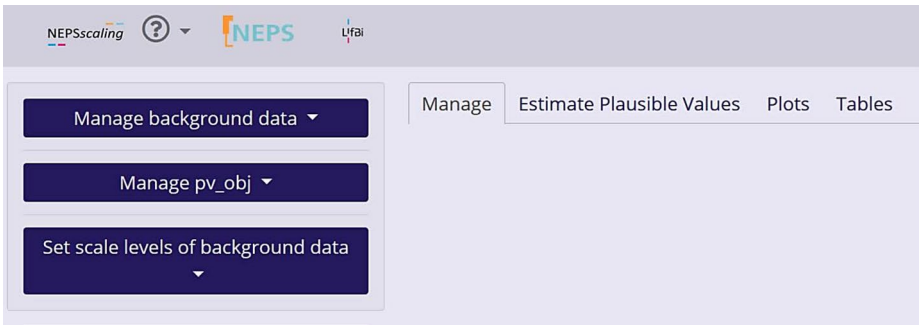


Fig. 1 Start screen of NEPS*scaling*

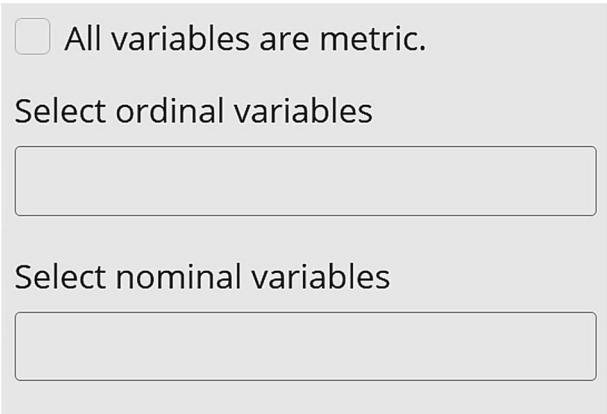


Fig. 2 Setting of scale level in the background data

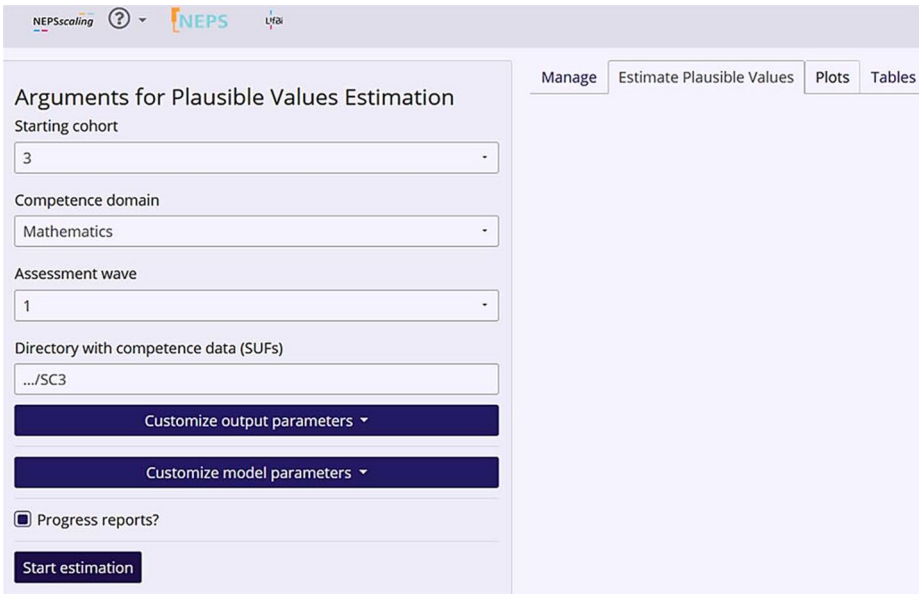


Fig. 3 Necessary input arguments for plausible values estimation

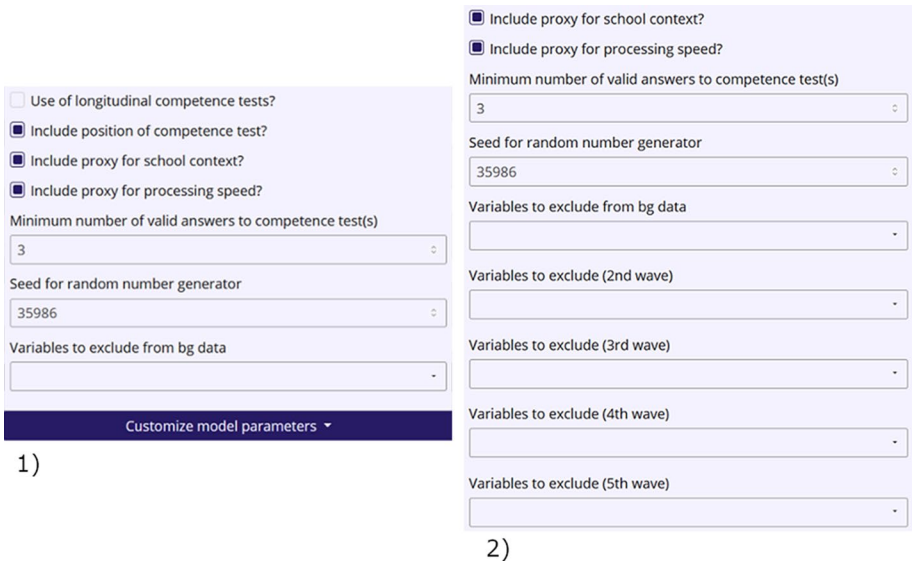


Fig. 4 Further parameters for tweaking the (1) cross-sectional and (2) longitudinal plausible values estimation

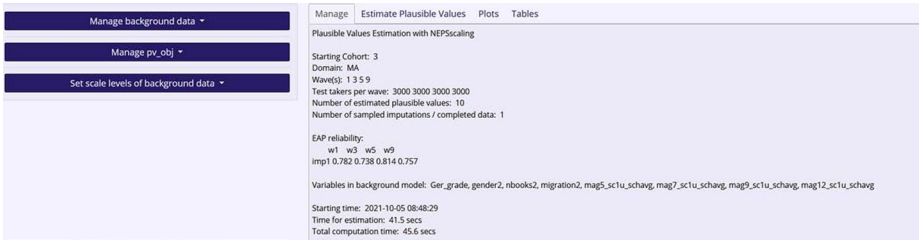


Fig. 5 Short summary of the plausible values object after estimation or import

Application 2: Longitudinal math competence in the 5th grader starting cohort

Using the Shiny app is less concise, but also more intuitive if there is little to no prior experience with R. The functions corresponding to application 1 are illustrated below; additional functionality is shown in the online supplemental material accompanying this paper. After the package has been installed, NEPS*shiny* can be launched by invoking the following R code in an R session started, for example, by RStudio.

```
NEPSshiny(launch.browser = TRUE)
```

The start screen of the app can be seen in Fig. 1. It allows the import and export of the underlying background data and previously estimated plausible values objects. It can be reached at any time by pressing the NEPS*scaling* logo in the upper left corner. To estimate a new set of PVs, the first step is to import the background data. Tabular data in R, SPSS and Stata file formats of up to 30 MB size can be imported. The data selection works by browsing the file system. The button "Remove background data" removes the

currently available object from the Shiny app's working environment. The inspection of background data is covered in the supplemental material (Additional file 1: Figs. S1–S3).

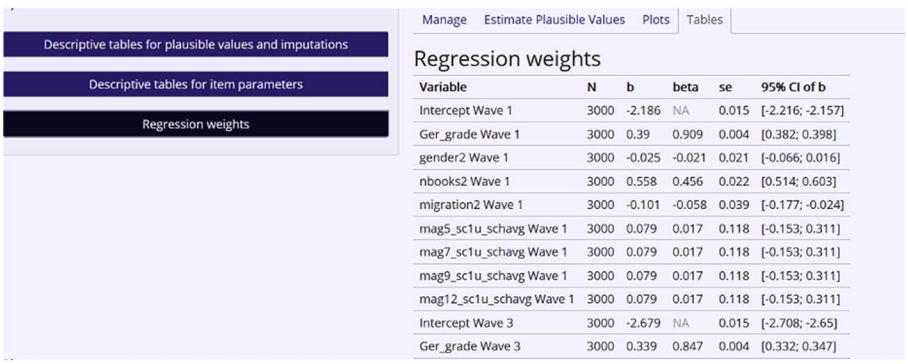
After uploading the background data, the scale level of the data needs to be set (see Fig. 2) because categorical data is processed differently than metric variables in the imputation and estimation steps of NEPS*scaling*. The differentiation of ordinal and nominal variables becomes important for the aggregation of the imputed background data.

Next, we enter the "Estimate Plausible Values" tab (see Fig. 3). The application example is concerned with estimating PVs for the 5th grader cohort, SC 3. The goal is to obtain mathematics PVs for longitudinal analyses. Figure 3 shows how the SC, competence domain and assessment wave have already been set. Please note that the assessment wave can be any of the waves for the SC and domain combination in longitudinal estimation. Furthermore, the path to the competence data, set to the current working directory by default, has also been changed to the current location of the SC 3 SUFs.

In this configuration, ten cross-sectional plausible values for wave 1 are estimated. To switch to longitudinal estimation, the button at the top of the expanded "Customize model parameters" field as seen in Fig. 4, subfigure 1, needs to be checked. This leads to the further expansion of the field seen in subfigure 2 of Fig. 4. It is now also possible to exclude variables of the background data from the estimation of plausible values for specific assessment waves.

If all parameters are set to the intended model, the "Start estimation" button (see Fig. 3) can be pressed and the PVs are estimated. A summary of the current plausible value object can be inspected in the "Manage" tab (corresponding to the print()

1)



2)

Fig. 6 Summary tables of (1) item and (2) regression parameters

statement; see Fig. 5) and in the "Tables" tab where the item parameters (subfigure 1 of Fig. 6) and the estimated regression weights (subfigure 2 of Fig. 6) are displayed. Further visual inspection of the object is possible and shown in the supplemental material.

Summary

As can be seen in the application examples above, the main benefits of NEPS*scaling* lie in its simplicity. With this package, NEPS data users can use PVs for their population level analyses in only a few steps and without worrying whether the unknown background model of the PVs available in scientific use files actually fits their own analyses. Nevertheless, there are further notices regarding the package.

Before starting any analysis, users are required to have substantial knowledge on their used data. The package NEPS*scaling* does not release the user from their duty of knowing and understanding the data. The use of custom background data means that this data has to be prepared additionally by the users. However, data has to be prepared for the analyses in any case and the analysis data is identical to the background data of the PVs in most cases. The added amount of time and effort, thus, reduces to considering additional variables for the imputation of missing values and the estimation model. Similarly, the measurement models are restricted to tested scaling models. If a more flexible IRT model is desired, for example the three-parameter logistic IRT model, users will have to resort to other software solutions such as the R package *TAM*, on which NEPS*scaling* is based, or *mirt* or Mplus. Further information on the original scalings of the tests in the NEPS are available in technical reports on the NEPS website. It is important to mention that competence data between different starting cohorts cannot be linked as the estimation of plausible values is only possible within a specific starting cohort. Furthermore, the package is not available via CRAN, it is downloadable from the NEPS RDC's website without any further requirements or restrictions.

The package will be updated after each new release of competence data in the SUFs so that the users can use PVs for NEPS competence assessments as soon as possible after the SUF release. NEPS*scaling* was specifically designed to conduct analyses with NEPS data, therefore, it is required that users have access to NEPS data. Researchers are required to sign a data use agreement with the NEPS Data Center for data access.

In conclusion, NEPS*scaling* provides PVs for all scalable competence measurements in the NEPS with the additional benefit of automatically implementing an imputation scheme for the background data. Because of the non-parametric nature of the CART algorithm, it only requires the selection of the correct variables for the imputation model, but not its full specification. Non-linear relationships in the data are implicitly considered in the imputation with CART. Furthermore, NEPS*scaling* makes estimating PVs easier than non-study-specific packages like mirt or TAM since it does not require the specification and testing of a scaling model by the user. The quality of the estimation is checked and tested by the maintainers specifically for each model. The graphical user interface also allows easy use by researchers not proficient in the statistical programming language R.

Abbreviations

CART	Classification and regression trees
GUI	Graphical user interface
IRT	Item response theory
LSAS	Large scale assessment study
MICE	Multiple imputation through chained equation
NAEP	National Assessment of Educational Progress
NEPS	National Educational Panel Study
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
PV	Plausible value
RDC	Research data center
SC	Starting cohort
SUF	Scientific Use File
TIMSS	Trends in International Mathematics and Science Study
WLE	Warm's weighted maximum likelihood estimate

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40536-022-00145-5.

Additional file 1. Electronic supplement. The electronic supplement is provided as a standard PDF file. It contains further screenshots of the graphical userinterface to illustrate its functionalities beyond the basic estimation of plausible values for NEPS data.

Acknowledgements

This paper uses data loosely modeled after the National Educational Panel Study (NEPS): Starting Cohort 3 (https://doi.org/10.5157/NEPS:SC3:10.0.0) and 6 (https://doi.org/10.5157/NEPS:SC6:12.0.1). From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

Author Contributions

AS analyzed and interpreted the data used in this study. AS drafted significant parts of the manuscript. EZ substantially revised the manuscript. All authors read and approved the final manuscript.

Funding

The package development was funded by the Leibniz Institute for Educational Trajectories.

Availability of data and materials

The package NEPSscaling, including the data used for the examples, can be found at https://www.neps-data.de/Data-Center/Overview-and-Assistance/Plausible-Values.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 17 February 2022 Accepted: 3 December 2022
Published online: 26 December 2022

References

Blossfeld, H. P., & von Maurice, J. (2011). Education as a lifelong process. *Zeitschrift für Erziehungswissenschaft, 14*(S2), 19–34. https://doi.org/10.1007/s11618-011-0179-2

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

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification and regression trees*. Routledge. https://doi.org/10.1201/9781315139470

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

Bürkner, P. C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06
Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*(4), 330–351.
Doove, L. L., van Buuren, S., & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics and Data Analysis, 72*, 92–104. https://doi.org/10.1016/j.csda.2013.10.025
Fischer, L., Gnambs, T., Rohm, T., & Carstensen, C. H. (2019). Longitudinal linking of Rasch-model-scaled competence tests in large-scale assessments: A comparison and evaluation of different linking methods and anchoring designs based on two tests on mathematical competence administered in grades 5 and 7. *Psychological Test and Assessment Modeling, 61*(1), 37–64.
Fuß, D., Gnambs, T., Lockl, K., & Attig, M. (2021). *Competence data in NEPS: Overview of measures and variable naming conventions (starting cohorts 1 to 6)*. Leibniz Institute for Educational Trajectories (LIfBi), National Educational Panel Study (NEPS).
Lechner, C. M., Bhaktha, N., Groskurth, K., & Bluemke, M. (2021). Why ability point estimates can be pointless: A primer on using skill measures from large-scale assessments in secondary analyses. *Measurement Instruments for the Social Sciences, 3*(1), 1–16. https://doi.org/10.1186/s42409-020-00020-5
Linacre, J. M. (1989). *Many-faceted Rasch measurement*. Chicago: MESA Press.
Little, R. J. A. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics, 6*(3), 287. https://doi.org/10.2307/1391878
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
Lüdtke, O., Robitzsch, A., & Grund, S. (2017). Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological Methods, 22*(1), 141–165. https://doi.org/10.1037/met0000096
Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174. https://doi.org/10.1007/BF02296272
Meng, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science, 9*(4), 538–558. https://doi.org/10.1214/ss/1177010269
Muthén, L. K., & Muthén, B. O. (1998). Mplus user's guide. *Muthén and Muthén, 61*, 290–300.
OECD. (2017). *Pisa 2015 technical report*. https://www.oecd.org/pisa/data/2015-technical-report/
Pohl, S., & Carstensen, C. H. (2013). Scaling of competence tests in the national educational panel study-many questions, some answers, and further challenges. *Journal for Educational Research Online, 5*(2), 189–216. https://doi.org/10.25656/01:8430
Raghunathan, T., Reiter, J., & Rubin, D. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics, 19*, 1–16.
Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Danish Institute for Educational Research
Robitzsch, A., Kiefer, T., & Wu, M. (2021). *Tam: Test analysis modules*. https://CRAN.R-project.org/package=TAM
Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys Wiley series in probability and statistics*. Wiley. https://doi.org/10.1002/9780470316696
Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147
Scharl, A., Carstensen, C. H., & Gnambs, T. (2020). *Estimating plausible values with NEPS data: An example using reading competence in starting cohort 6*. NEPS Working Papers/Survey Papers.
van Buuren, S. (2012). *Flexible imputation of missing data*. Chapman and Hall/CRC. https://doi.org/10.1201/b11826
van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation, 76*(12), 1049–1064. https://doi.org/10.1080/10629360600810434
von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful. *IERI Monograph Series, 2*, 9–36.
Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427–450. https://doi.org/10.1007/BF02294627
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
Weis, M., & Reiss, K. (2019). Pisa 2018—Ziele und inhalte der studie. In K. Reiss, M. Weis, E. Klieme, & O. Köller (Eds.), *PISA 2018: Grundbildung im internationalen Vergleich* (pp. 13–20). London: Waxmann.
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

Publisher's Note