NEPSscaling: Plausible Value Estimation for Competence Tests Administered in the German National Educational Panel Study

Anna Scharl
Leibniz Institute for Educational Trajectories: Leibniz-Institut fur Bildungsverlaufe
https://orcid.org/0000-0003-0081-1893
Eva Zink (eva.zink@lifbi.de)
Leibniz Institute for Educational Trajectories at the University of Bamberg: Leibniz-Institut fur Bildungsverlaufe https://orcid.org/0000-0002-3671-6372

Research Article

Keywords: National Educational Panel Study, plausible values, competence, NEPSscaling, large-scale assessment

Posted Date: May 13th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1360912/v1

RESEARCH

NEPSscaling: Plausible Value Estimation for Competence Tests Administered in the German National Educational Panel Study

Author1 and Author2*

*Correspondence: email2
Affiliation, City, Country
Full list of author information is available at the end of the article

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 only requires the user to prepare the data for the background model. Then, the appropriate item response model including the linking approach adopted for the NEPS is automatically selected, while a nested multiple imputation scheme handles missing values in the background data. For novice users, a graphical interface is provided that does not require 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

1 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 and 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 Germany and follows newborns to pensioners in repeated assessments across their life courses (Blossfeld and 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 antecedents and outcomes of education in Germany. In LSAs, competencies are typically analyzed as plausible values (PVs) because these allow for unbiased population parameter estimation at the population level (Lüdtke and Robitzsch, 2017; Lechner et al, 2021). Precise PV estimation requires the specification of a background model that is appropriate for the research question at hand. 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. A challenge of longitudinal LSAs such as the NEPS is their growing data base because 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 and 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 and Reiter, 2010; Weirich et al, 2014) and estimates the appropriate item response models following the scaling standards in the NEPS (Pohl and 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 plausible values and then describe the basic functionality of NEPS*scaling*. The use of the package is demonstrated in two examples that show how to estimate plausible values using R syntax or the graphical user interface.

2 Background

2.1 Plausible Values

Following the NEPS*scaling* standards (Pohl and 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

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

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 plausible values technique is an extension of IRT models via a latent regression of the person parameters on background variables (Lüdtke and 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

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

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 variables $\boldsymbol{x_i}$ for person i

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

with $\beta_L = (\beta_1, ..., \beta_L)^T$ 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 interaction or non-linear terms that are part of the planned analyses (Bondarenko and Raghunathan, 2016; Meng, 1994). Similarly, it may be sensible to add further variables to improve the imputation of the background data.

This also highlights that plausible values 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).

[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.

2.2 Classification and Regression Trees

The missing data strategy in LSAs for plausible values estimation typically encompasses re-defining missing values as an additional dummy variable during the recoding of the background data. This cannot be seen as effectively handling missing data (Lüdtke et al, 2017; Schafer and Graham, 2002) 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.

We chose the CART algorithm to impute the background data (Burgette and 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 and 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 is that the splitting of child nodes automatically implies non-linear relationships in the data without having to explicitly model them.

3 About NEPSscaling

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, information about the implemented competences tests and assessment waves, and exporting PVs for different statistical software. NEPS*scaling* is also available as a Shiny app which can be launched by invoking

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

3.1 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, while the latter reports known deviations in comparison to the point estimates (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): Vs 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 for all waves of the specified cohort should be estimated.
- *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 school average point estimate (WLE) of the competence should be 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.

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_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).

3.2 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 background data for the PV estimation needs to be prepared by the user to ensure congeniality with the intended statistical analyses. This data preparation can be done using any statistical software. The only requirement is that the resulting background data is stored in tabular format either in SPSS, Stata, or R's rds format (when using the graphical user interface) or it is imported into the current R session as a *data.frame*

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

(when invoking the estimation using the R script). Importantly, missing data in the background data must be coded as R's *NA* values and categorical variables have to be converted to R's *factors*. 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 NEPSshiny 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.

3.3 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-Assistance/Plausible-Values. Previous package versions and example code that shows how to use the package to estimate PVs in different cohorts are also available.

4 Applications

In the following, two example applications are presented that use simulated data sets included in the package. The data was modeled after 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 exampled uses the NEPS*scaling* Shiny app. The input data in both cases 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 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 plausible values. 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. Simulated examples for background data are included in the package; real examples can be found in further user examples given at the download site of NEPS*scaling*.

4.1 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, there are three steps until PVs are ready for further processing. The first step consists of loading all necessary packages for importing the background data into R and loading NEPS*scaling*. Then, PVs can be estimated.

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

The argument *path* assumes that the SUFs for this example reside in the subfolder *SC6* of the current R working directory. 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 plausible values, the fixed item difficulties, and the estimated latent regression weights.

```
## 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).

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

4.2 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.

– Insert Figure 1 around here –

The start screen of the app can be seen in Figure 1. It allows the import and export of the underlying background data and previously estimated pv_obj objects. It can be reached at any time by clicking 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 (Figures S1 to S3).

– Insert Figure 2 around here –

After uploading the background data, the scale level of the data needs to be set (see Figure 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 Figure 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.

– Insert Figure 3 around here –

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 Figure 4, subfigure 1, needs to be checked. This leads to the further expansion of the field seen in subfigure 2 of Figure 4. It is now also possible to exclude variables of the background data from the estimation of plausible values for specific assessment waves.

– Insert Figure 4 around here –

If all parameters are set to the intended model, the "Start estimation" button (see Figure 3) can be clicked and the PVs are estimated. A summary of the current pv_obj can be inspected in the "Manage" tab (corresponding to the print() statement; see Figure 5) and in the "Tables" tab where the item parameters (subfigure 1 of Figure 6) and the estimated regression weights (subfigure 2 of Figure 6) are displayed. Further visual inspection of the object is possible and shown in the supplemental material.

– Insert Figure 5 and 6 around here –

5 Summary

As can be seen in the application examples above, the main benefits of NEPSscaling lie in its simplicity. With this package, NEPS data users can use plausible values 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.

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, users will have to resort to other software solutions and information on the original scalings of the tests available in technical reports on the NEPS website. Furthermore, the package is not available via CRAN, but can be downloaded via the NEPS RDC's website.

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.

In conclusion, NEPSscaling 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 does not require the specification of an imputation model, but implicitly considers non-linear relationships in the data. 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.

Acknowledgements

This paper uses data loosely modeled after the National Educational Panel Study (NEPS): Starting Cohort 3 (http://dx.doi.org/10.5157/NEPS:SC3:10.0.0) and 6 (http://dx.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.

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

Abbreviations
CART – Classification and regression trees
GUI – Graphical user interface
IRT – Item response theory
LSAS – Large scale assessment study
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

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

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
A1 analyzed and interpreted the data used in this study. A1 drafted significant parts of the manuscript. A2 substantially revised the manuscript. All authors read and approved the final manuscript.

Authors' information
Author1 https://orcid.org/0000-0000-0000-0000
Author2 https://orcid.org/0000-0000-0000-0000

Author details
Affiliation, City, Country.

References
Blossfeld HP, von Maurice J (2011) Education as a lifelong process 14:19–34,
Bondarenko I, Raghunathan T (2016) Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models 35(17):3007–3020,
Bürkner PC (2017) brms: An R package for Bayesian multilevel models using Stan 80(1):1–28,
Burgette LF, Reiter JP (2010) Multiple imputation for missing data via sequential regression trees 172(9):1070–1076,
Chalmers RP (2012) mirt : A multidimensional item response theory package for the R environment 48(6):1–29,
Doove L, Van Buuren S, Dusseldorp E (2014) Recursive partitioning for missing data imputation in the presence of interaction effects 72:92–104,
Fischer L, Gnambs T, Rohm T, Carstensen CH (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 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). URL https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/Kompetenzen/Overview_NEPS_Competence-Data.pdf
Lüdtke O, Robitzsch A (2017) Eine Einführung in die Plausible-Values-Technik für die psychologische Forschung 63(3):193–205,
Lüdtke O, Robitzsch A, Grund S (2017) Multiple imputation of missing data in multilevel designs: A comparison of different strategies. 22(1):141–165,
Lechner CM, 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 3(1):1–16,
Linacre JM (1989) Many-faceted Rasch measurement
Masters GN (1982) A Rasch model for partial credit scoring 47(2):149–174,
Meng XL (1994) Multiple-imputation inferences with uncongenial sources of input 9(4):538–558,
Muthén LK, Muthén BO (1998) Mplus User's Guide. Eighth Edition, Muthén & Muthén
Pohl S, Carstensen CH (2013) Scaling of competence tests in the National Educational Panel Study - many questions, some answers, and further challenges 5(2):189–216, , publisher: Waxmann
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. URL https://CRAN.R-project.org/package=TAM
Rubin DB (1987) Multiple Imputation for Nonresponse in Surveys. Wiley Series in Probability and Statistics, John Wiley & Sons, Inc.,
Schafer JL, Graham JW (2002) Missing data: Our view of the state of the art. 7(2):147–177,
Warm TA (1989) Weighted likelihood estimation of ability in item response theory 54(3):427–450,

Weirich S, Haag N, Hecht M, Böhme K, Siegle T, Lüdtke O (2014) Nested multiple imputation in large-scale assessments 2(1):9,
Weis M, Reiss K (2019) PISA 2018 – Ziele und Inhalte der Studie. In: Reiss K, Weis M, Klieme E, Köller O (eds) PISA 2018: Grundbildung im internationalen Vergleich, Waxmann, pp 13–20

Figures

Figure 1: Start screen of NEPS*scaling*

Figure 2: Setting of scale level in the background data

Figure 3: Necessary input arguments for plausible values estimation

Figure 4: Further parameters for tweaking the 1) cross-sectional and 2) longitudinal plausible values estimation

Figure 5: Short summary of *pv_obj* after estimation or import

Figure 6: Summary tables of 1) item and 2) regression parameters

Additional Files

Additional file 1 — Electronic supplement

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

Figures

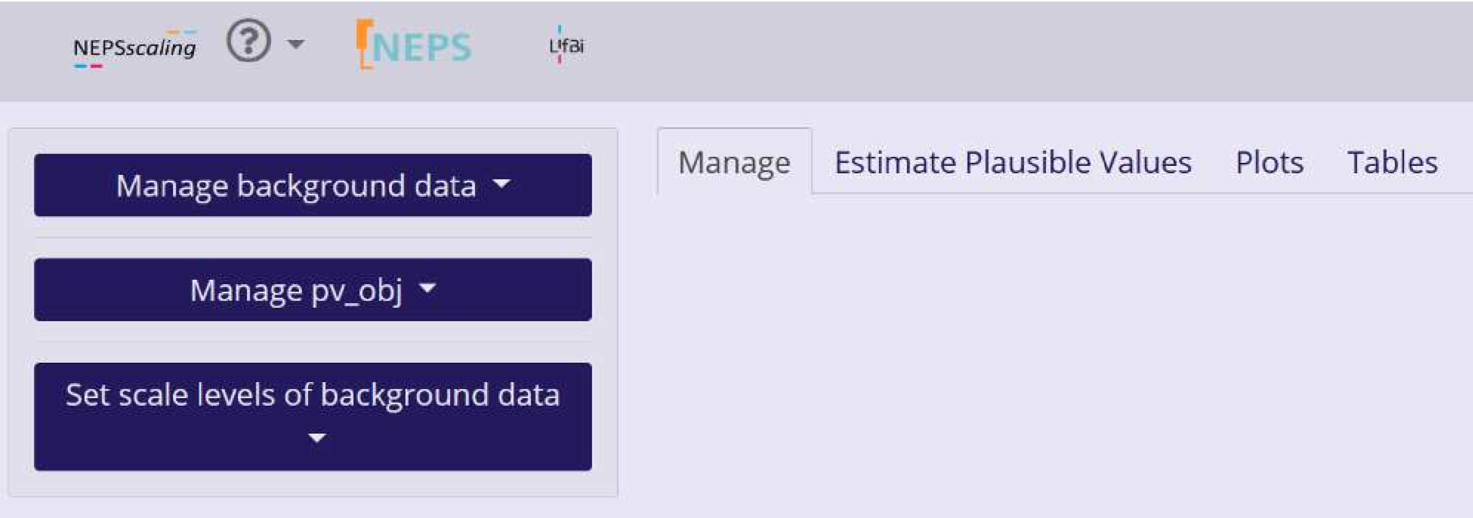

Figure 1

Start screen of NEPSscaling

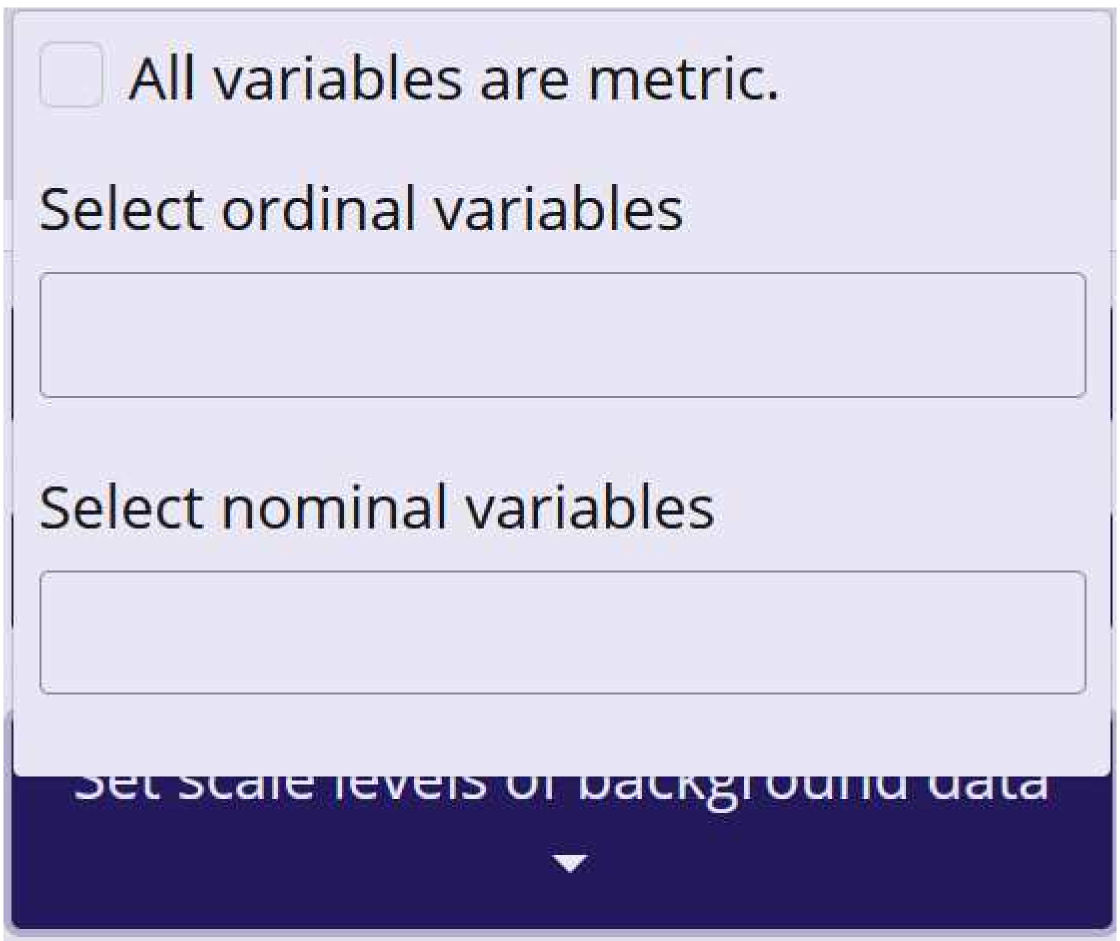

Figure 2

Setting of scale level in the background data

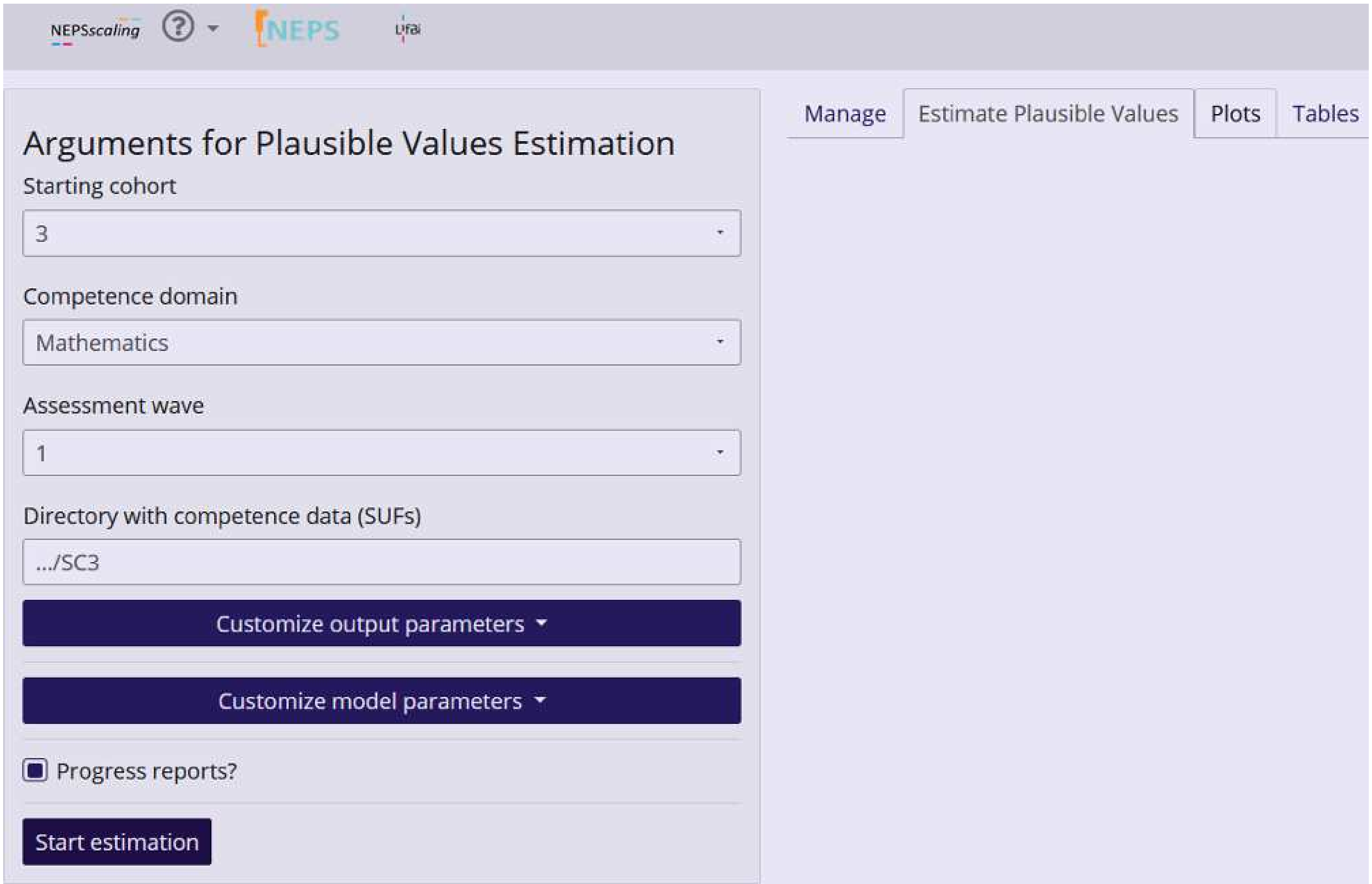

Figure 3

Necessary input arguments for plausible values estimation

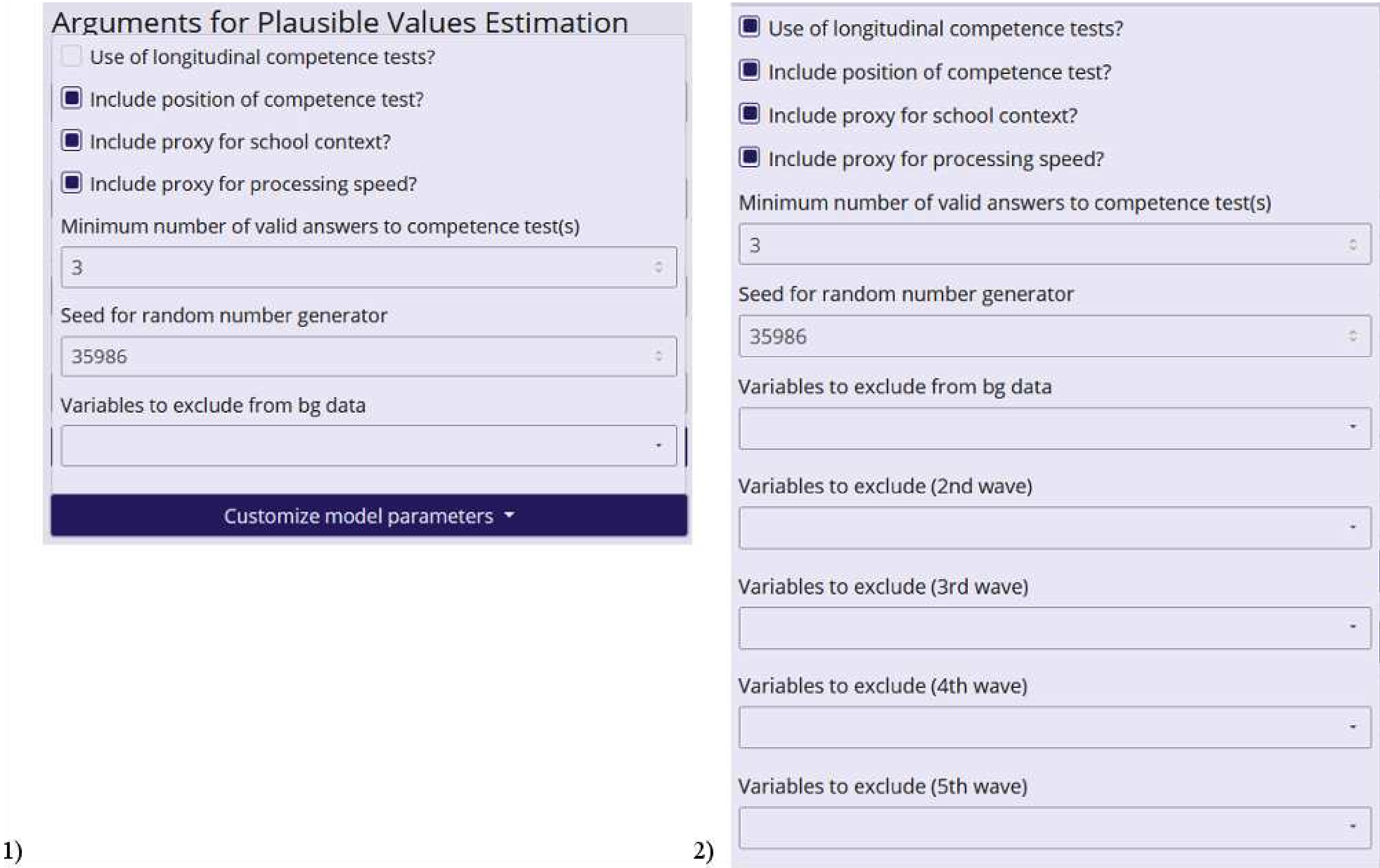

Figure 4

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

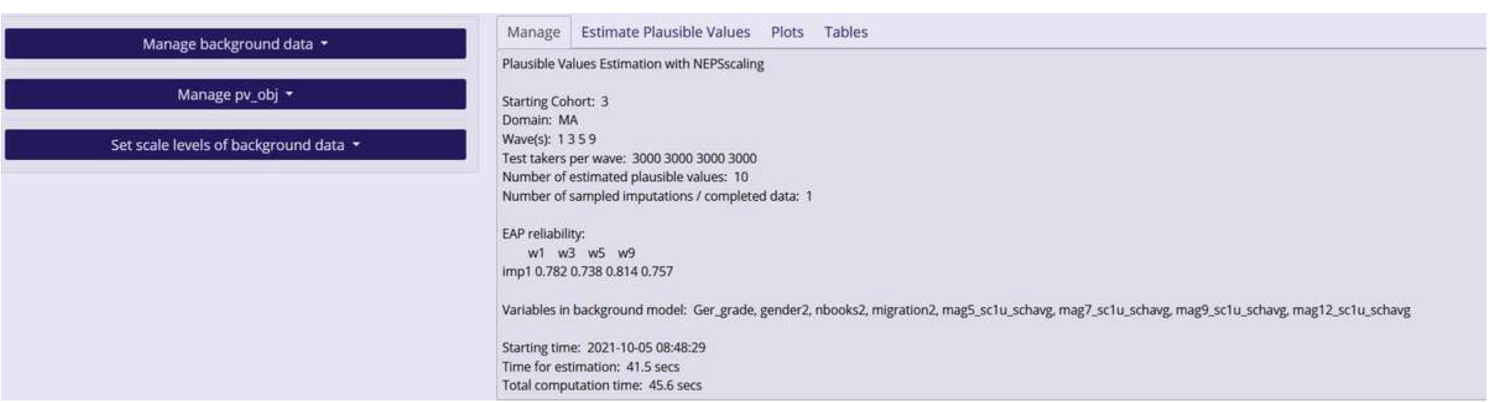

Figure 5

Short summary of *pv_obj* after estimation or import

Figure 6

Summary tables of 1) item and 2) regression parameters

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- nepsscalingsupplement.pdf