NEPS
National Educational Panel Study

NEPS Working Papers

Steffi Pohl & Claus H. Carstensen

NEPS Technical Report – Scaling the Data of the Competence Tests

NEPS Working Paper No. 14

Bamberg, October 2012

SPONSORED BY THE
Federal Ministry of Education and Research

Working Papers of the German National Educational Panel Study (NEPS)
at the University of Bamberg

The NEPS Working Papers publish articles, expertises, and findings related to the German National Educational Panel Study (NEPS).
The NEPS Working Papers are edited by a board of researchers representing the wide range of disciplines covered by NEPS. The series started in 2011.

Papers appear in this series as work in progress and may also appear elsewhere. They often represent preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character.

Any opinions expressed in this series are those of the author(s) and not those of the NEPS consortium.

The NEPS Working Papers are available at
http://www.uni-bamberg.de/neps/publikationen/neps-working-papers/

Editorial Board:
Jutta Allmendinger, WZB Berlin
Cordula Artelt, University of Bamberg
Jürgen Baumert, MPIB Berlin
Hans-Peter Blossfeld, EUI Florence
Wilfried Bos, University of Dortmund
Edith Braun, HIS Hannover
Claus H. Carstensen, University of Bamberg
Henriette Engelhardt-Wölfler, University of Bamberg
Johannes Giesecke, University of Bamberg
Frank Kalter, University of Mannheim
Corinna Kleinert, IAB Nürnberg
Eckhard Klieme, DIPF Frankfurt
Cornelia Kristen, University of Bamberg
Wolfgang Ludwig-Mayerhofer, University of Siegen
Thomas Martens, DIPF Frankfurt
Manfred Prenzel, TU Munich
Susanne Rässler, University of Bamberg
Marc Rittberger, DIPF Frankfurt
Hans-Günther Roßbach, University of Bamberg
Hildegard Schaeper, HIS Hannover
Thorsten Schneider, University of Leipzig
Heike Solga, WZB Berlin
Petra Stanat, IQB Berlin
Volker Stocké, University of Kassel
Olaf Struck, University of Bamberg
Ulrich Trautwein, University of Tübingen
Jutta von Maurice, University of Bamberg
Sabine Weinert, University of Bamberg

Contact: German National Educational Panel Study (NEPS) − University of Bamberg − 96045 Bamberg − Germany − contact.neps@uni-bamberg.de

NEPS Technical Report – Scaling the Data of the Competence Tests

Steffi Pohl & Claus H. Carstensen, Otto-Friedrich-Universität Bamberg, National Educational Panel Study

E-Mail-Adresse des Erstautors:

steffi.pohl@uni-bamberg.de

Bibliographische Angaben:

Pohl, S. & Carstensen, C. H. (2012): NEPS Technical Report – Scaling the Data of the Competence Tests (NEPS Working Paper No. 14). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

We thank Kerstin Haberkorn for her assistance in developing scaling standards and Carmen Köhler and Katinka Hardt for valuable feedback on previous versions of this paper. We also thank Natalie Boonyaprasop for English proofreading.

NEPS Technical Report – Scaling the Data of the Competence Tests

Abstract

The National Educational Panel Study (NEPS) aims at investigating the development of competences across the whole life span. Tests for assessing the different competences are developed in NEPS and response data is collected from study participants on different competence domains in different age cohorts. The data of the competence tests are scaled using models of Item Response Theory (IRT). In the Scientific Use File (SUF) competence data are provided for researcher in form of item responses, manifest scale scores, as well as plausible values that allow investigating latent relationships. This paper aims at achieving different purposes. First, at describing the scaling model used to estimate competence scores in NEPS. This includes aspects like dealing with different response formats and accounting for missing responses in the estimation, as well as describing the parameters that are estimated in the model. Second, describing the various analyses that are performed for checking the quality of the competence tests. This includes item fit measures, differential item functioning, test targeting, unidimensionality, and local item independence. And third, outlining different approaches on how the competence data provided in the SUF may be used for further analyses. While the sections on the scaling model and the quality check are written for researchers familiar with IRT, the section on how to use the competence scores provided in the SUF is written for substantive researchers interested in using competence scores to investigate research questions. ConQuest-syntax is provided for some analyses examples.

Keywords

Item Response Theory, Scaling, Competence Tests, Technical Report, Plausible Values, Partial Credit Model

Content

1. Introduction 4
2. Competence tests in NEPS 5
3. General scaling model 6
3.1 The Item Response Model 6
3.2 Incorporating different response formats 7
3.3 Treating missing responses in items 8
3.4 Estimation of competence scores 9
3.4.1 Weighted maximum likelihood estimates 9
3.4.2 Plausible values 9
4. Checking the quality of the test 10
4.1 Fit of subtasks 10
4.2 Item fit 11
4.3 Differential item functioning 12
4.4 Test targeting 13
4.5 Dimensionality of the test 13
4.6 Local item dependence 14
5. How to work with competence data 14
5.1 Using weighted maximum likelihood estimates 16
5.2 Using plausible values provided in the Scientific Use File 17
5.3 Including the measurement model in the analysis 20
5.4 Estimating plausible values 21
References 23
Appendix 29
Appendix A: Response formats in the competence test 29
Appendix B: ConQuest-Syntax and output for investigating the effect of gender on ability 30
Appendix C: ConQuest-Syntax for estimating plausible values 31

1. Introduction

The National Educational Panel Study (NEPS) aims at investigating the development of competences across the whole life span. Tests for assessing the different competences are developed in the NEPS and response data is collected from study participants in different age cohorts. In this manuscript we describe the scaling model, the analyses performed for quality checks of the test items, as well as how competence scores in NEPS may be used in further analyses.

Please note that the scaling procedure described in this paper was primarily developed for scaling competence tests in the domains of reading competence, mathematical competence, scientific literacy, and information and communication technologies literacy. Nevertheless, the scaling procedure described here serves as a guideline for the scaling of other competence tests in NEPS. Consequently, the scaling of most of the NEPS competence data followed the scaling models and procedures described here. Deviations from these guidelines may be indicated for particular test instruments and will be reported in the technical reports of the respective competence domains. Also note that established test instruments are deployed in NEPS as well. These are scaled according to the instructions in the test handbook.

Many of the competence tests used in NEPS have been constructed specifically for implementation in NEPS. There are different stages of test construction. A large set of items is first developed and tested in different preliminary and pilot studies. The results of data analyses from these preliminary studies are then used for item selection and optimization. In the main studies only items that showed good psychometric properties in preliminary studies are used. Thus, in main studies we assume the items to have a good item fit. Nevertheless, the quality of the test is checked in the main studies to ensure data quality. The scaling model and quality analyses described in this paper are those performed for the main study data. Note, however, that the model and most of the analyses were also used to analyze the data of preliminary studies.

This paper aims at achieving different purposes. First, describing the scaling model used to estimate competence scores. Second, describing the analyses performed for checking the quality of the tests developed in NEPS. And third, outlining how the competence data provided in the Scientific Use File (SUF[1]) may be used for further analyses. Note that while the present paper focuses on describing the analyses models, the respective results of the quality checks of the tests are presented in the technical reports of the different competence domains and age cohorts. While the sections on the scaling model and the scaling procedures are written for researchers familiar with Item Response Theory (IRT), the section on how to use the competence scores provided in the Scientific Use File (SUF) is written for substantive researchers interested in using competence scores to investigate their research questions.

[1] Information on the access to data of the SUF as well as further information can be found on https://www.neps-data.de/de-de/datenzentrum.aspx

2. Competence tests in NEPS

In the National Educational Panel Study a variety of competence domains are assessed. Consequently, various tests are constructed to adequately measure the respective competence domains in different age cohorts. Life-span domains, such as reading competence, mathematical competence and scientific literacy, are assessed coherently across the life span in order to allow for investigating change over time. A framework has been developed for each competence domain and tests are constructed according to this framework for different age cohorts (see e.g., Gehrer et al., 2012; Hahn et al., 2012; Neumann et al., 2012). Other competence domains are specific for certain age groups (e.g., orthography, Frahm et al., 2011) or subpopulations (e.g., native language Russian or Turkish, Kristen et al., 2011). Many tests that are implemented in NEPS have been constructed specifically for this purpose and most of them are scaled based on Item Response Theory. Additionally, already established tests are deployed in NEPS (e.g., for measuring vocabulary; Berendes, Weinert, Zimmermann, & Artelt, 2012). These are scaled according to the scoring procedure described in the respective handbook. An overview of the competence tests used in the National Educational Panel Study is given by Weinert et al. (2011).

The tests assessing life-span competence domains are constructed according to a theoretical framework. With regard to reading competence, for example, the framework consists of text functions (informational, commenting, literary, instructional, and advertising texts) and cognitive requirements (finding information in text, drawing text related conclusions, and reflecting and assessing) into which the items may be classified (Gehrer et al., 2012). Regarding mathematics (Neumann et al., 2012), there are different content areas (quantity, change and relationships, space and shape, and data and chance) and cognitive components (mathematical communicating, mathematical arguing, modeling, using representational forms, mathematical problem solving, and applying technical abilities and skills). The framework for scientific literacy (Hahn et al., 2012) distinguishes between the knowledge of basic scientific concepts and facts (knowledge of science) and an understanding of scientific processes (knowledge about science). In information and communication technologies (ICT) literacy the framework (Senkbeil et al., 2012) includes both information literacy and technological literacy. This framework additionally distinguishes between different process components (define, access, manage, create, integrate, evaluate, and communicate) and content areas (word processing, tables, presentations, e-mail, and internet). When constructing the competence tests, items are developed that represent all different aspects of the framework.

In the life-span competence tests, there are four different response formats[2]. These are a) simple multiple choice, b) complex multiple choice, c) matching items, and d) short-constructed responses. Examples of the different response formats are given in Appendix A. In simple multiple choice (MC) items there are four response options with one option being correct and three response options functioning as distractors (i.e., they are incorrect). Complex multiple choice (CMC) items consist of a number of subtasks with one correct

[2] Note that the response formats described here are those used in the IRT-scaled competence tests that are constructed to measure life-span domains (i.e., reading, mathematical, and scientific literacy). Further response formats may occur in other competence tests in NEPS. How these are dealt with is described in the respective technical reports.

answer out of two response options. Matching (MA) items require the test taker to match a number of responses to a given set of statements. MA items are only used for measuring reading competence and usually require assigning headings to paragraphs of a text. Some MA items consist of as many responses as there are statements, while others contain more response options than there are statements. Short-constructed response (SCR) items are only used in mathematics tests. They usually require a response in form of a number.

In most of the studies in NEPS two competence domains are assessed together within the same wave. Usually reading and mathematics are assessed together in the first wave, while scientific literacy and ICT literacy are assessed together in the second wave. When different tests are jointly administered, the order in which the tests are administered must be considered. Previous studies (Adams & Carstensen, 2002) have shown that the position of item blocks within one booklet does have an effect on item difficulty. In order to account for effects of the position of a test within a booklet, the order of the competence domains was rotated. In the first wave, participants were randomly assigned to either work on the reading test first and then on the mathematics test – or vice versa[3]. The test order is then fixed and will be continued in subsequent waves. Subjects who received the reading test first in the first wave will always work on the reading test before the mathematical test in the following waves.

3. General scaling model

In this chapter we will introduce the general scaling model used to scale the competence data in NEPS. We will specifically point out how different response formats and missing responses were modeled and which test scores are estimated for the SUF.

3.1 The Item Response Model

The NEPS competence tests are constructed according to domain-specific frameworks (Weinert et al., 2011; Gehrer et al., 2012; Neumann et al., 2012; Hahn et al., 2012; Senkbeil et al., 2012). For every competence domain a unidimensional construct is assumed and the item construction is guided by several test characteristics, such as content or cognitive requirements. For each test, the number of items from the different conceptual differentiations is deliberately chosen to represent their relevance for the domain. Thus, an item response model for scaling NEPS competence data should preserve the weighting implemented by the number of items defined for each combination of conceptual differentiations (see Pohl & Carstensen, 2012). Consequently, the Rasch model (Rasch, 1960/1980) or extensions of the Rasch model (i.e., the partial credit model; Masters, 1982) were chosen as scaling models since they preserve the weighting of items by construction. Furthermore, compared to the 2PL model (Birnbaum, 1968) as a plausible alternative IRT model, the interpretation of the Rasch model scale values is more straightforward (Wilson, 2003, 2005).

For scaling the competence data, the Mixed Coefficients Multinomial Logit Model (Adams, Wilson, & Wang, 1997), which is implemented in the software ConQuest (Wu, Adams,

[3] Note that there are exceptions from this rule in some starting cohorts.

Wilson, & Haldane, 2007), was used. The Mixed Coefficients Multinomial Logit Model is a general model that includes the Rasch and the partial credit model as a special case.

Item parameter estimates are obtained using marginal maximum likelihood estimation incorporating the EM algorithm. In this estimation the ability distribution of the persons are assumed to be normal (Adams, Wilson, & Wang, 1997). Moderate deviations from the assumed ability distribution do not impact the quality of the item parameter estimates (Mislevy, 1984). Person scores on the latent competence domains are estimated in two different approaches. The first approach consists of computing weighted maximum likelihood estimates (WLEs; Warm, 1989), which are point estimates and best represent each participant's competence regarding observed responses only (and the assumed item response model). As a second approach for providing competence scores, the common latent distributions of competence scores and selected context variables will be provided in form of plausible values. This approach allows for unbiased population level analyses of competence distributions and context variables, and it requires a rather complex conditioning model. The two approaches of estimating competence scores are described in chapter 3.4.

As test forms were presented to the test takers in different positions within the booklet (see section 2) position effects may be present in the estimated competence scores. When estimating item and person parameters, individual test scores are corrected for the position effect in the estimation of ability scores. This correction follows the rationale of the correction of booklet effects in the Program for International Student Assessment (PISA; see, e.g., OECD 2005), where booklet effects were accounted for in form of main effects.

3.2 Incorporating different response formats

As described above, different response formats are present in the competence tests. These are simple multiple choice items, complex multiple choice items, matching items, and short-constructed responses. Items with simple multiple choice responses as well as short-constructed responses result in dichotomous variables which are analyzed using the Rasch model. Complex multiple choice items and matching items consist of item bundles with a common stimulus and challenge the assumption of local stochastic independence of the single responses (see, e.g., Yen, 1993). In order to account for this dependence, the items may be aggregated to an ordinal score for each item bundle and may be analyzed via the partial credit model (e.g., Andrich, 1985; Zhang, Shen, & Cannady, 2010).

The partial credit model (Masters, 1982) assumes the probability to respond in the higher of two adjacent categories to follow the same item characteristic curve as in the Rasch model for dichotomous data. Assuming the response probabilities over all possible item scores to be one, the following model for ordinal data can be derived:

$$P(X_{in} = x \mid \theta) = \frac{\exp\left[\sum_{k=0}^{x}(\theta_n - \delta_{ik})\right]}{\sum_{h=0}^{K_i}\exp\left[\sum_{k=0}^{h}(\theta_n - \delta_{ik})\right]} \qquad \text{(Eqn. 1)}$$

with

$$\sum_{k=0}^{0}(\theta-\delta_{ik})\equiv 0,$$

where X_{in} denotes the response of person n on item i, x denotes the categories of the item, θ_n denotes the trait level of person n, and δ_{ik} the item parameters.

When an instrument is composed of items with different response formats, the question of how to weight the different item response formats in the item response model arises. In the Rasch and the partial credit model, the weight of an item is determined by its maximum score. For polytomous items the maximum number of score points usually corresponds to the number of score options, while for dichotomous items the maximum number of score points credited for each item is one. These score points have the function of item discrimination constants. The maximum score of an item is, however, to some extent arbitrarily chosen. Should a correct response to a dichotomous subtask of a CMC item be scored as high as a correct response on a simple MC item with four response options? Furthermore, CMC items with a higher number of score points might be more informative and more discriminating than complex items with a lower number of subtasks. In order to determine an appropriate weight for the different response formats, the impact of different approaches to scoring items with different response formats on discrimination and the fit of the items has been investigated in empirical studies on NEPS competence data (Haberkorn, Pohl, Carstensen, & Wiegand, 2012). The authors concluded that – given that dichotomous items (MC and SCR items) are scored with zero (for an incorrect response) and one (for a correct response) – scoring each subtask of a CMC and MA item with 0.5 points best fits the empirical data. Therefore, in the scaling of competence data in NEPS, the general rule was posed to credit each score point of a polytomous item with 0.5 points in the response model. With this scoring rule a simple MC item (with four response options) is weighted twice as much as a single subtask of a CMC item (with two response options) and CMC and MA items with more subtasks are weighted higher than those with less subtasks. Exceptions from this rule may be applied when there are theoretical reasons for an alternative scoring of the item.

3.3 Treating missing responses in items

There are four different kinds of missing responses in the competence data. These are 1) items that are not administered (due to the testing design), 2) invalid responses (e.g., more than one response to a simple MC item), 3) omitted items, and 4) items that were not reached due to time limits (i.e., omitted responses after the last valid response). The ignorability of the missing responses depends on the causes of missingness. While in most test designs missing responses due to not administered items are missing completely at random, the omitted and not reached items are usually nonignorable (Mislevy & Wu, 1988) and often depend on the difficulty of the item and on the ability of the person. There are different approaches for treating missing responses, and several studies (e.g., Culbertson, 2011; De Ayyala, Plake, & Impara, 2001; Finch, 2008; Lord, 1974; Ludlow and O'leary, 1999; or Rose, von Davier, & Xu, 2010) have investigated their performance. These studies showed that ignoring missing responses, multiple imputation (Rubin, 1987), as well as model-based

approaches (Glas & Pimentel, 2008; Holman & Glas, 2005; O'Muircheartaigh & Moustaki, 1999) result in unbiased parameter estimates, whereas treating missing responses as incorrect for either item or person parameter estimation results in biased parameter estimates. Pohl, Gräfe, and Hardt (2011), Gräfe (2012) as well as Pohl, Gräfe, and Rose (2012) compared the different approaches for treating missing responses in different domains and cohorts in NEPS and found indications that ignoring missing responses in the scaling model results in unbiased item and person parameter estimates. This closely resembles the results found in simulation studies (Rose, von Davier, & Xu, 2010). For scaling the competence data in NEPS, all kinds of missing responses were thus ignored.

3.4 Estimation of competence scores

Within the Scientific Use Files, two different estimates of competences will be provided. Weighted maximum likelihood estimates (WLEs) will be available with the first data release. Due to the complex generation process, plausible values will be made available in later updates of data releases.

3.4.1 Weighted maximum likelihood estimates

The WLE as a typical point estimate expresses the most likely competence score for each single person given the item responses of that person. This is what would be reported back to single participants, since point estimates are unbiased estimates of individual scores. WLE are corrected for a part of the bias of the maximum likelihood estimate (MLE), which tends to have too extreme values and leads to overestimation of the variance of the ability distribution (Warm, 1989). In many cases the variance of WLE is still larger than the variance of the true person parameters. However, with sufficiently large numbers of items their variance gets close to the latent variance of the response model (Walter, 2005). WLE scores include measurement error components and their variance includes error variance. An inference on this uncertainty of measurement can, however, not be drawn from the WLEs, since the variance of the true parameters cannot be separated from the error variance. In order to obtain unbiased analysis results of population parameters that are purified from measurement error, latent modeling of common distributions of competences and context variables using plausible values may be employed.

3.4.2 Plausible values

Plausible values are basically multiple imputations (Rubin, 1987) for the latent variable in an item response model (Mislevy, 1991). They are Bayesian scale scores obtained as random draws from a posteriori competence distribution that is a function of the item responses and the context variables included in a multivariate response model. Aggregating the random draws on a group level shall give unbiased group level results. This approach requires the inclusion of the context variables used in later analyses into the measurement model. These context variables are included in the model via latent regression or multidimensional IRT-models and are used for the estimation of the abilities of the persons (see, e.g., Adams, Wu, & Carstensen, 2007; Carstensen, Knoll, Rost, & Prenzel, 2004). The plausible values then reflect relations between context variables and competence measures, the uncertainty of these relations due to measurement error, and the measurement error in general.

Since plausible values are simply random draws from the posterior distributions, and any one set of plausible values will give unbiased estimates of population and subpopulation

level competence distributions, these values are *not* suitable "scores" for the individuals in the sample. The average of these estimates within subgroups will, however, give unbiased estimates of the group-level statistics of interest. In general, five sets of plausible values are drawn (recommendation by Little & Rubin, 1987), although also more can be drawn.

Since in NEPS many context variables are collected, the challenge in estimating plausible values is to find an appropriate conditioning model that is suitable for a variability of possible research questions. Since this is a very complex endeavor, plausible values will be provided in later releases of the SUF.

4. Checking the quality of the test

Before estimating ability scores, the quality of the items in each competence domain and for each cohort is checked in various analyses. These include fit of the subtasks of CMC and MA items, item fit, differential item functioning, test targeting, unidimensionality, and local item dependence.

4.1 Fit of subtasks

Before aggregating the subtasks of the complex items to a polytomous score, in a first step, we test the fit of the subtasks of all CMC and MA items. For this purpose, a dichotomous Rasch model is fitted to all disaggregated items (for each CMC and MA item the single subtasks are included in the analysis[4]). The weighted mean square (WMNSQ; Wright & Masters, 1982), its t-value, the empirically approximated item characteristic curve (ICC), and the point biserial correlation of the responses with the total score (relative number of correct responses on the total number of valid responses) are used to evaluate the fit of the single subtasks.

Subtasks with a dissatisfying fit are excluded from the following analyses. Subtasks with a satisfactory fit are aggregated to polytomous variables indicating the number of correct responses in a given CMC or MA item. For CMC and MA items with many subtasks, the frequency of a total score of 0 on the polytomous variable representing the respective CMC or MA item is often relatively low and will most likely result in estimation problems when included in an IRT analysis. Therefore, categories with an absolute frequency of less than N = 200 were subsumed with the adjacent category.

For MA items, in which the number of responses equals the number of statements to be matched to, there is an extremely high linear dependency of the responses to the subtasks. If a person correctly matches four out of five responses to the five statements, the fifth response is perfectly determined. If persons get four out of five subtasks right, a correct response to the last subtask is determined. As a consequence, a score of one point less than the maximum score does not occur (or very rarely). In order to avoid estimation problems and since there is no additional information gained from this last response, the second last category is collapsed with the maximum score category for these variables. Note that, due to the collapsing of categories, the score of the polytomous CMC and MA items does not necessarily indicate the number of correctly answered subtasks.

[4] ignoring the local item dependence of these items

4.2 Item fit

The fit of the items is evaluated using various fit measures. The quality of all items is evaluated by the weighted mean square (WMNSQ; Wright & Masters, 1982; Wu, 1997), correlations of the item score with the total score, the estimated item discrimination, and the item characteristic curve. Additionally, a distractor analysis is performed (correlation of incorrect response with the total score). The MNSQ and WMNSQ (Wright & Masters, 1982) are indices that describe the deviation of the observed probability for a correct response from the model implied probability for a given ability level. In contrast to the MNSQ, the WMNSQ weights deviations from the curves more strongly for response probabilities close to 0.5 and less for very low and very high probabilities. A WMNSQ near 1 indicates a good fit. A WMNSQ lower than 1 indicates an overfit, that is, the item discriminates more than assumed in the model. WMNSQ scores greater than 1 usually occur if the discrimination of the item is low or if the empirically estimated response curve is not monotone increasing. WMNSQ below 1 are considered to be a less serious violation to model fit than WMNSQ greater than 1. The respective t-values are inference statistical measures for the null hypothesis that the WMNSQ equals one. Note that, due to the large sample size, most of the t-values indicate a significant deviation from a WMNSQ of 1. As Wu (1997) showed, the fit statistics depend on the sample size. We applied the rules of thumb displayed in Table 1 for evaluating the WMNSQ and its respective t-value in NEPS.

Table 1: Rules of thumb for WMNSQ fit values

	Noticable item misfit		**Considerable item misfit**	
N	Weighted MNSQ	t	Weighted MNSQ	t
7500	> 1.15	>| 6|	> 1.20	>|8|
15000	> 1.10	>| 8|	> 1.15	>|10|

In addition to the WMNSQ and its t-value, the correlations of the item score with the total score are evaluated. The correlation of the item score with the total score should be positive. Subjects with a high ability should be more likely to score high on the item than subjects with a low ability. Furthermore, correlations of incorrect response options and the total score are evaluated. The correlations of the incorrect responses with the total score allow for a thorough investigation of the performance of the distractors. A good item fit would imply a negative or a zero correlation of the distractor with the total score. Distractors with a high positive correlation may be an indication of an ambiguity in relation to the correct response. Based on experiences with several data sets in NEPS and other large scale studies, we formulated the rules of thumb displayed in Table 2 for evaluating the respective correlations in NEPS.

Aside from fit statistics, which provide a form of aggregated information, we investigate the fit of the items by comparing the model-implied item characteristic curve with the empirically estimated one. If considerable deviations of the curves were found, the appropriateness of the item was further investigated by the item developers.

Table 2: Rules of thumb for the correlation of item score and distractor with the total score

	Good	Acceptable	Problematic
Item score	> 0.30	> 0.20	< 0.20
Distractor	< 0.00	< 0.05	> 0.05

Although the scaling model assumes Rasch-homogeneity, that is, it equally weights the items that have the same response format, the equality of the item discriminations is empirically tested. In order to do this, item discriminations are estimated using a generalized partial credit model (2PL; Muraki, 1982). Besides evaluating the estimated discriminations on item level, the 2PL model is compared to the 1PL model using Akaike's (1974) information criterion (AIC) as well as the Bayesian information criterion (BIC, Schwarz, 1978).

4.3 Differential item functioning

Differential item functioning (DIF) is a form of testing measurement invariance across subgroups. DIF exists when subjects with the same trait level have a different probability of endorsing an item. In such cases, the item functions differ between different subgroups and the item favors one of the subgroups. In such cases, one should not compare the competence of the different subgroups to each other, since differences in competence scores may be due to differences between subgroups and differences in item difficulties. NEPS aims at constructing tests that are fair for subgroups. Therefore, DIF is one exclusion criterion in the process of item development, and it is investigated for the NEPS competence data of the main studies for different subgroups.

Differential item functioning is investigated in NEPS by applying multiple-group IRT analyses in which difficulties are allowed to vary across subgroups. In the model applied, main effects of the group variable as well as interaction effects of item and group are modeled. Note that main effects do not indicate DIF but rather mean differences in competence between subgroups. DIF is, however, present when there are differences in item difficulties. Based on experiences with preliminary data, we consider absolute differences in estimated difficulties that are greater than 1 logit as very strong DIF, absolute differences between 0.6 and 1 noteworthy for further investigation, differences between 0.4 and 0.6 as considerable but not sincerely, and differences smaller than 0.4 as no considerable DIF. Furthermore, an overall test is performed by comparing a model including item by group interactions to a model that only allows for main effects of the group variable. AIC and BIC are used to evaluate model fit.

The variables for which DIF is tested in NEPS include gender, migration background, the number of books at home (as a proxy for socioeconomic status), school type or degree, as well as test position in the booklet. Further DIF variables were used when considered important for a cohort or competence domain. Migration background was chosen as a DIF variable since NEPS specifically oversamples persons with a migration background and aims at drawing conclusions about the educational development of persons with a migration background. To specify the migration background for the persons, the native country of the target person and of the parents was used. A dichotomous variable was constructed indicating no migration background when both the target person him- or herself as well as

his or her parents were born in Germany. A target person was classified as having a migration background if either the person itself or one of its parents was born in a foreign country. This means that persons of the first, 1.5, and second generation were assigned to have a migration background (see Stanat & Segeritz, 2009, for details of the indicators). In case the relevant variables were not available in some cohorts, a proxy (e.g., the first language of the target person and the parents) was used. The number of books at home was used as a proxy for socioeconomic status (see, e.g., Baumert & Bos, 2000; Elley, 1994). A dichotomous variable was constructed distinguishing between target persons with up to 100 books and target persons with more than 100 books at home (see, e.g., Paulus, 2009). This distinction was used, since it results in commensurate group sizes. School type was included in DIF analyses as a dichotomous variable differentiating between persons in grammar school (German: Gymnasium) or having an A-level degree (German: Abitur), and persons in other school types or having lower school degrees. A dichotomous distinction was used because the concept of grammar school is similar in all Federal States, while other school forms are not coherently defined in the different Federal States. DIF for the position of the test in the booklet was estimated in order to ensure that test order does not distinctively affect certain items. When there were more than 300 cases with missing responses to the DIF variable, the persons with missing responses were included as a separate group in the DIF analysis. Small group sizes may cause estimation problems when estimating DIF (Clauser & Mazor, 1998; Ziesky, 1993). Thus, when a DIF variable contained less than 300 missing responses, the respective cases were deleted from the DIF analysis.

4.4 Test targeting

The information of an item, and thus, the measurement precision, is highest for subjects that have an ability similar in size to the difficulty of the item. The test information is the sum of the item information and depends on the distribution of the item difficulties. The measurement precision of an ability estimate is described by its standard error. The standard error of measurement is inversely related to the test information. A good test targeting is obtained if there is high test information, that is, a low standard error, for the whole range of the ability distribution. Since ability distributions are often normally distributed, items are constructed so that their difficulties resemble this distribution. This results in very precise ability estimates for (the many) subjects with average ability and in less precise ability estimates for (the few) subjects with a very low or very high ability[5].

4.5 Dimensionality of the test

All competence tests in NEPS are constructed to measure a unidimensional construct. Nevertheless, the dimensionality of the test is evaluated in the data. Subdimensions are postulated based on the different aspects of test construction (Neumann et al., 2012; Gehrer et al., 2012; Hahn et al., 2012; Senkbeil et al., 2012).

The subdimensions differ for different competence domains. They are, however, all based on construction principles for the test. For reading (Gehrer et al., 2012), for example, subdimensions are based on a) cognitive requirements and b) text functions.

[5] Note that this is different in adaptive testing designs, in which an equally high measurement precision of ability is aimed for the whole range of the ability distribution.

Multidimensional models are estimated using quadrature (for low dimensional models) or Monte Carlo (for high dimensional models). The dimensionality is evaluated by comparing model fit of multidimensional models to the unidimensional one as well as by estimated correlations between subdimensions. If the correlations between the subdimensions are high (>.95, see Carstensen, in press), an unidimensional model may be assumed. Model comparison is performed based on information criteria, that is, on Akaike's (1974) information criterion (AIC) as well as the Bayesian information criterion (BIC; Schwarz, 1978). The BIC takes the number of estimated parameters into account and, thus, prevents from overparametrization of models.

4.6 Local item dependence

Local item dependence (LID) may occur for item bundles (Rosenbaum, 1988) or testlets (Wainer & Kiely, 1987) that share a common stimulus. This is especially the case for reading tests where a set of items refers to the same text. Local item dependence may lead to an underestimation of standard errors, bias in item difficulty estimates, inflated item discrimination estimates, overestimation of the precision of examinee scores, and overestimation of test reliability and test information (Sireci, Thissen, & Wainer, 1991; Wainer, 1995; Wainer & Lukhele, 1997; Wainer & Thissen, 1996; Wainer & Wang, 2000; Yen, 1993). We investigated the impact of LID using either the Rasch testlet model (Wang & Wilson, 2005) or a multidimensional model. In the multidimensional model, a separate dimension is modeled for each item bundle for which LID is expected, and the correlation between the dimensions is used as a measure of LID. In the Rasch testlet model, the variance of the items is additively decomposed into a common part and a testlet specific part, allowing us to investigate the variance proportion of the item responses that emerges due to the item bundle. If applicable, local item dependence is evaluated in a test using either of the models.

5. How to work with competence data

There are different ways of working with competence data in substantive analyses. In the Scientific Use File, item data as well as two different competence scores (WLEs and plausible values) are or will be provided. Depending on the specific research question, whether measurement error should be accounted for, and on the knowledge about specifying measurement models, different approaches can be recommended for users. We will focus on four different approaches. In the following, we will describe how to choose between these approaches. Note, that these four approaches are not an exhaustive list. They are a selection of approaches considered useful for working with competence data provided in NEPS.

In a first step, a researcher working with competence data has to decide whether he or she wants to account for measurement error in the analysis. Measurement error may have an impact on the point and interval estimates of population parameter (e.g., means, variances, correlations, or regression coefficients). Measurement error masks relationships (e.g., Aiken & West, 1991; Cohen, Cohen, West, & Aiken, 2003; Maddala, 1977) and may result in biased parameter estimates (e.g., correlations). This is called attenuation bias. If, for example, the true correlation between two competence scores measured with tests like in NEPS is 0.7, latent correlations will reveal this correlation, while the correlation estimated using manifest competence scores is confounded with measurement error and, therefore, lower (in the

simulated example here: cor = 0.61). Also, percentiles and population variance estimates are biased when using manifest competence scores and unbiased when using latent competence scores. Both manifest and latent competence scores provide unbiased point estimates of population means and regression coefficients. However, standard errors of these estimates are underestimated using manifest scores. An illustrative example of these properties can be found in OECD (2009b). For an example in large scale assessment studies see Adams, Wu, and Carstensen (2007).

Using manifest competence scores is a very convenient approach since competence scores are included in the analyses just as any other manifest variables. Manifest competence scores in NEPS are provided in form of weighted maximum likelihood estimates (WLEs). Working with WLEs is described in chapter 5.1. Accounting for measurement error in the analyses does require more complex procedures. We describe three different approaches accounting for measurement error in an analysis with competence data that allow for latent variable modeling. One of these approaches is to use plausible values which are provided in later releases of the Scientific Use File. Chapter 5.2 describes how to work with these data. Plausible values are, however, not provided in the first releases of the SUF, and even then they may not be convenient for all research questions[6].

As described above, for using plausible values to investigate specific research questions, the respective variables used in these research questions need to be included in the conditioning model. If, for example, a researcher wants to investigate the effect of gender on mathematical competence, gender must be included in the conditioning part of the measurement model generating the plausible values. If the interaction of gender and self-concept on mathematical competence shall be investigated, the interaction of gender and self-concept needs to be included in the measurement model. A researcher needs to be aware of this and check whether the variables she/he would like to use in the analysis are included in the measurement model (in the respective functional form) for the plausible values. If the variables of interest are not part of the conditioning model[7], one might want to consider using one of the other two approaches for investigating latent relationships with competence scores.

If a researcher wants to investigate latent relationships with competence data from the first data releases, he or she may either include the measurement model in their analyses (approach 3, chapter 5.3) or estimate plausible values him- or herself (approach 4, chapter 5.4). While approach 3 is very elegant, since it incorporates the measurement and the analysis in one model, it may easily become a quite complex task if many variables are included in the structural model. Estimating plausible values yourself is a good alternative, when one wants to estimate latent relationships in more complex analyses. How to choose between the different approaches is depicted by a decision tree in Figure 1.

[6] With the release of plausible values in later releases of the Scientific Use File, a documentation of the estimation model will be provided that allows judging whether the plausible values provided are suitable for answering a specific research question.

[7] Note that not all variables used to answer a research question necessarily need to be in the conditioning model. It is sufficient that the relevant variables are well explained by the variables that are in the conditioning model. Research is currently conducted investigating the robustness of analyses results to misspecification of the conditioning model given that a large set of context variables is included in the conditioning model.

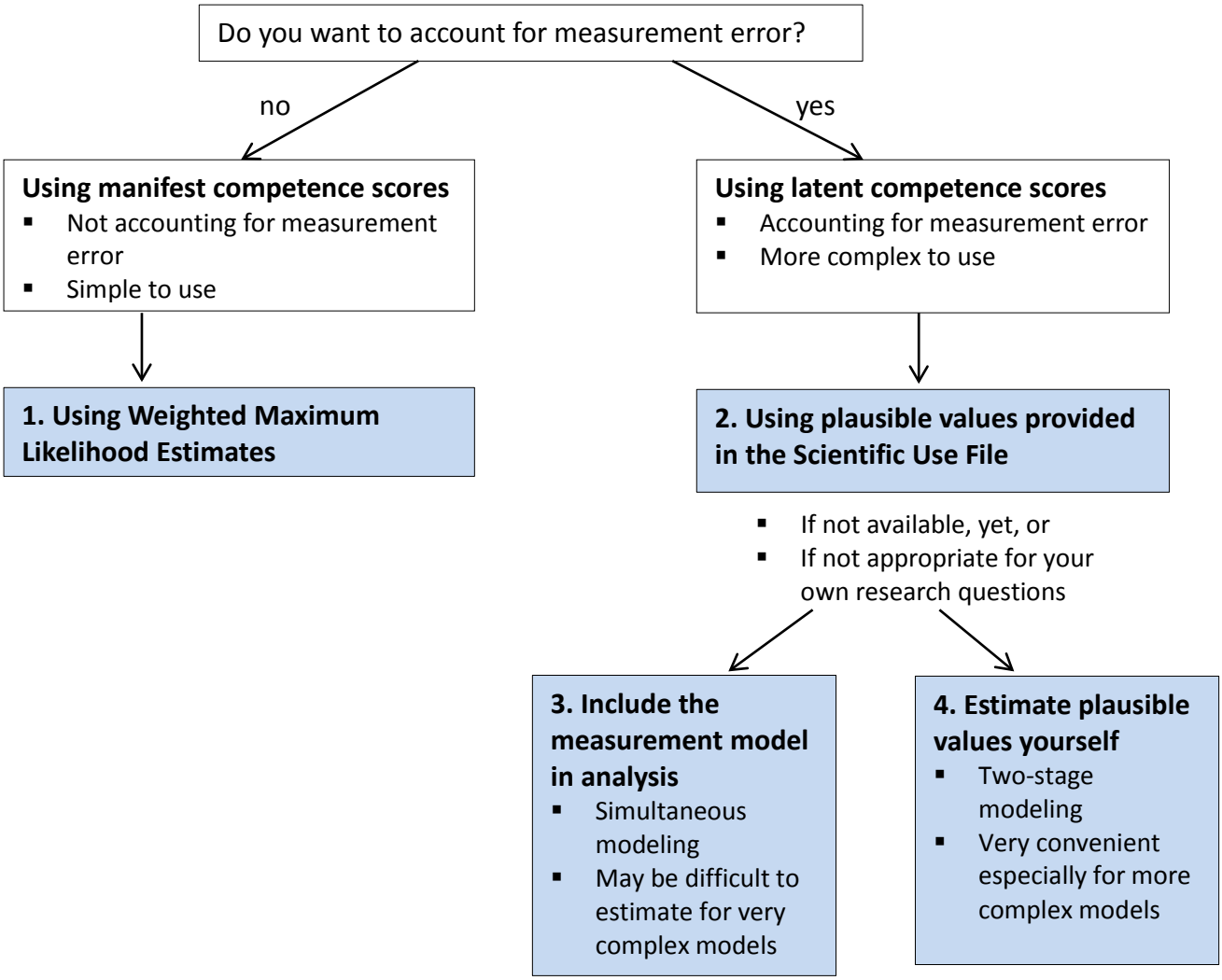

Figure 1: Decision tree for choosing an approach for working with competence data

In the following sections the different approaches are described in more detail. Note that when performing substantive analyses, sampling weights also have to be taken into account (see Aßmann et al., 2011, for a description of the sampling weights in NEPS).

5.1 Using weighted maximum likelihood estimates

The WLEs are available in the first release of the competence data. As described above, the weighted maximum likelihood estimates are the best point estimates of the individual competence scores. WLEs have very similar properties as sum scores of correct responses in test items. Both are manifest scores that do not take measurement error into account. WLEs have, however, some advantages over sum scores. They do, for example, facilitate adequate treatment of missing responses (see section 3.3 for the treatment of missing responses) and comparability of competence scores over different waves and cohorts. WLEs may be used in further analyses just as other manifest variables. WLEs in the first SUF release are constrained to have a mean of zero. Values above zero indicate abilities above average, while WLE scores below zero indicate abilities below average. The variance of the WLE scores is not restricted. Note that WLE scores are confounded by measurement error. If latent distributions or relations to other variables are to be estimated, one of the other three methods for working with competence data are advised for use.

5.2 Using plausible values provided in the Scientific Use File

Plausible values are competence estimates that describe a range of possible values for a student's ability. They are ability estimates that take information about group membership (e.g., gender or socioeconomic status) into account. For each individual a distribution of possible values is estimated that is derived from his or her responses to the items, as well as from information about the person (conditioning variables such as gender or socioeconomic status). Plausible values are random draws from this distribution. Thus, if a male student, say Ben, and a female student, say Laura, have the same response pattern to the items in the competence test, they will not have the same distribution of possible ability values. This is because information on conditioning variables (e.g., gender) is taken into account and the ability estimates are shifted towards the mean of the competences of the respective group. So, if males have on average a higher mathematical competence than females, Ben will have higher plausible values than Laura, although both answered the same items correctly. This example emphasizes that plausible values are *not* suitable for estimating individual ability scores. They do *not* represent unbiased scores on the individual level! They are, therefore, not suitable for reporting back ability levels to students. Plausible values aim at estimating unbiased group level statistics (e.g., mean differences in competence scores between males and females). They do provide unbiased point estimates of group statistics (e.g., means, mean differences, variances, correlations, regression coefficients) and of their corresponding standard errors (and therefore p-values).

Usually, five plausible values are provided for each person and each competence domain. Analyzing plausible values requires performing as many analyses as there are plausible values for each individual. The analysis is performed repeatedly with each set of plausible values. The mean of the statistic (e.g., a mean difference or a regression coefficient) over all analyses gives the population estimate for that statistic. The variability of the estimated statistic across the five analyses reflects the uncertainty of the estimate that is due to measurement error. The estimated standard error of the statistic reflects uncertainty due to sampling. Combining the variance due to measurement error and the variance due to other sources, like sampling error, gives adequate estimates of the standard error of the group level statistic (Rubin, 1987).

Rubin (1987) and Mislevy (1991) provide formulas for calculating the point estimate of the statistic of interest as well as for the appropriate standard error of this estimate. From the *m* analyses on the *m* datasets, an estimate of the statistic of interest is estimated as

$$\bar{Q}_m = \sum_{l=1}^{m} Q_{*l} \,/\, m \qquad \text{(Eqn.2)}$$

with Q being the statistic of interest, $l=1,\ldots,m$ being the sets of plausible values, and Q_{*l} being the statistic of interest estimated in data set l.

The associated variance of $\bar{Q}_m$ is

$$T_m = U_m + \frac{m+1}{m} B_m \qquad \text{(Eqn.3)}$$

where

$$U_m = \frac{1}{m}\sum_{l=1}^{m} U_{*l} \qquad \text{(Eqn.4)}$$

is the within-imputation variability (with U_{*l} being the variance estimate of the statistical parameter of interest in analysis *l*), and

$$B_m = \frac{1}{m-1}\sum_{l=1}^{m}(Q_{*l} - \bar{Q}_m)^2 \qquad \text{(Eqn.5)}$$

is the between-imputation variability. The α%-confidence interval for *Q* is

$$CI(Q) = Q \pm t_v(1-\alpha)\cdot T_m^{\frac{1}{2}} \qquad \text{(Eqn.6)}$$

with $t_v(1-\alpha)$ being the $1-\alpha$ percentile of the *t*-distribution with *v* degrees of freedom

$$v = (m-1)\left(1 + \frac{U_m}{\frac{m+1}{m}B_m}\right)^2 . \qquad \text{(Eqn.7)}$$

Table 3: Results of a hypothetical analysis. The table shows the point estimates, standard errors, and error variances of the mean difference in mathematical competence between males and females using each set of plausible values.

Analysis	**Estimate of mean difference**	**S.E.**	**Error variance (S.E.^{2})**
1	0.35	0.040	0.00160
2	0.33	0.035	0.00123
3	0.40	0.042	0.00176
4	0.32	0.037	0.00137
5	0.32	0.041	0.00168
Mean	0.344		0.00153
Variance	0.00113[8]		

[8] The variance was calculated according to formula (Eqn.5) for the estimation of the between-imputation variability

The following example serves to illustrate the estimation of point estimates and confidence intervals based on plausible values. As an example, we consider the research question of whether a gender difference exists in the mean mathematical ability score. Suppose $m = 5$, that is there are five plausible values for each individual on mathematical competence. Thus, if the mean difference between males and females is to be estimated, the mean difference needs to be estimated five times. In each analysis, the mean differences and the standard error are estimated based on one of the five sets of plausible values using any convenient software. Table 3 shows the results of our hypothetical analysis. For each set of plausible values there is a point estimate of the mean difference as well as a standard error (S.E.). The error variance displayed in Table 3 is the square of the standard error.

The point estimate of the mean difference is $\overline{Q}_m$ =0.344, the mean of the estimated mean differences in each of the five analyses. The within-imputation variability is 0.00153 (see Equation 4), the mean of the error variance estimates in the five analyses. This variation reflects sampling error. The between-imputation variability is 0.00133 (see Equation 5), which is the variance of the estimated mean differences in the five analyses. This variation reflects the measurement error. The associated variance T_m of the mean difference of 0.344 may now be calculated as 0.00153+[(5+1)/5]0.00113=0.002886 (see Equation 3). The respective standard error is then 0.0537215, the square root of this variance. The degrees of freedom for constructing a confidence interval are

$$15 \approx (5-1)\left(1+\frac{0.00153}{\frac{5+1}{5}0.00133}\right)^2$$

(see Equation 7) and, thus, the t-value of a two-sided test on an alpha-level of 5% is 2.131 (value obtained from the t-distribution). With the standard error, confidence intervals (CI) may be estimated and inference statistic performed. The 95% CI is

$$[0.344-2.131\cdot 0.0537215, 0.344-2.131\cdot 0.0537215]=[0.230, 0.459]$$

(see Equation 6). Since the confidence interval does not include zero, this mean difference significantly differs from zero on an alpha-level of 5%.

Of course, such analyses may not only be performed with mean differences, but with any statistical parameter of interest. Many statistical software programs (e.g., M*plus*, Muthén & Muthén, 1998-2010; Stata, McDonald, 2008/2011; or HLM, Raudenbush, Bryk, & Congdon, 2004) can deal with data sets using plausible values, and they do provide correct estimates of the parameter estimates and their standard errors. Illustrative data analysis manuals for working with plausible values in SAS (OECD, 2009a) and SPSS (OECD, 2009b) are provided for PISA data.

An advantage of plausible values provided in the Scientific Use File is that analyses can be performed just as with any other manifest variables and no specific software is required. The difference is that the analysis is performed several times (each with a new set of plausible values) and that the results of these analyses need to be combined using straightforward formulas. Thus, since analyzing available plausible values requires far less specific knowledge

than specifying the appropriate measurement model, plausible values are a tool, which enable a wide range of researchers to perform unbiased analyses.

If only one set of plausible values (instead of all five) is used for an analysis, the results will still give unbiased estimates of the statistic on the group level. The standard error of the statistic, however, will be underestimated and inferences will be too progressive. Note that the uncertainty due to measurement error is only incorporated if a number of analyses with different plausible values are performed and the results of the analyses are aggregated. Note that aggregating plausible values already on the individual level is not an appropriate analysis strategy, since the mean of plausible values is neither a good point estimate nor does it provide unbiased group-level estimates (e.g., von Davier, Gonzalez, & Mislevy, 2009).

Since finding an appropriate measurement model for drawing plausible values that are suitable for many research questions is very complex for the NEPS data, further research on appropriate models is necessary, and plausible values will be provided in later releases of the Scientific Use Files.

5.3 Including the measurement model in the analysis

The measurement model for a competence domain may also be included in structural analyses. This approach models measurement and structural part of the model simultaneously. It is a common approach in structural equation modeling, where both measurement and structural model are combined in one model. Similarly, this may also be done with IRT measurement models (Wilson & deBoeck, 2004). An advantage of this procedure is that latent relationships may be investigated (vs. manifest ones in the first approach using WLEs) and that standard errors may be estimated directly since everything is estimated in one model (vs. the two-step procedure encountered in the plausible-values approaches). This approach is also very convenient for dealing with missing values in the context variables, since the latent ability may be used for a simultaneous modeling of the missing responses (see Aßmann, Carstensen, Gaasch, & Pohl, 2012). This procedure may, however, reach its limitations when the number of latent variables becomes very high or the model becomes very complex (e.g., many latent variables, hierarchical structures). In these cases, estimation of the model may be difficult.

As an example, we included explanatory variables for the competence score in the measurement model in ConQuest (see Figure 2 for the graphical display of the model and the respective syntax in Appendix B[9]). In the example in Appendix B, gender (0 – female, 1 – male) is included in the model as a predictor for reading competence[10]. The regression coefficient estimated in the model (see ConQuest output in Appendix B) is a measure for the relationship between the latent ability score and gender. For identification the additive constant of the regression is set to zero. The unstandardized regression coefficient for gender is -0.175 logits, indicating that males have a 0.175 lower reading ability than females.

[9] Note that in this example there are only dichotomous items. In most of the competence data in NEPS, there will also be polytomous scored items. The respective scaling syntax will be provided in the technical reports of the different competence domains.

[10] Note that the position of the test in the booklet is also included in the model in order to account for order effects (see section 3.1)

This difference is statistically significant (SE=0.032, 95%-confidence interval [-0.238,-0.112]). The error variance is 1.36. Of course, also other parameters of interest (e.g., multiple regression coefficients) may be estimated and one may use any type of software that is capable of estimating IRT models (specifically the partial credit model) to include the measurement model in their analyses.

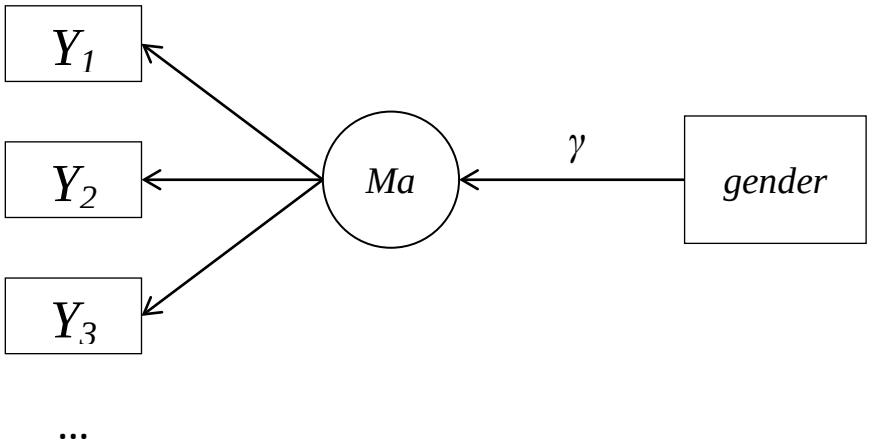

Figure 2: Including the measurement model of mathematical competence in the analysis of gender effects on mathematics.Y_i denote the response variables of the mathematics items and Ma the latent mathematics competence score.

5.4 Estimating plausible values

As has been described above, to investigate the relationship of competence measures with other variables, the relevant variables need to be included in the conditioning model for estimating plausible values. If this is not the case, one option is to estimate plausible values yourself, including the relevant variables in the conditioning model. Using plausible values is a two-step approach. In the first step, plausible values are estimated including relevant conditioning variables in the measurement model. In a second step, analyses of interest are performed using the different sets of plausible values and point and interval estimates are obtained as described in 5.2.

The advantage of this procedure is that the conditioning model needs to include only the variables of the particular research question. In contrast to including the measurement model in the analysis (see approach 5.3) this approach may easily deal with a larger number of conditioning variables. After estimating plausible values, classical statistical analyses may be performed in the usual way with any type of software. The analyses are, however, performed for each of the sets of plausible values (usually five), and the results of the analyses are aggregated as described in chapter 5.2. Note that, if nonlinear relationships are of interest, nonlinear terms need to be included in the conditioning model.

An example of a ConQuest-syntax for estimating plausible values for mathematical competence can be found in Appendix C. In this model, gender, school type, and age are included as main effects in the conditioning model. The resulting plausible values are suitable for analyses including these variables (or variables that are well explained by these variables). For example, one may investigate the following research questions:

1. Is there a mean difference in mathematical competence between males and females?
2. How much variance does school type explain on mathematical competence?
3. How well does school type explain mathematical competence, controlling for age?

The estimated plausible values in this example are not necessarily suitable for answering research questions like:

4. How much variance of mathematical competence do gender and socioeconomic status explain?
5. Is the effect of gender on mathematical competence moderated by age?

For answering research question 4, socioeconomic status needs to be included in the conditioning model. For answering research question 5, the interaction of gender and age must be considered in the conditioning model.

Note that not all variables used in the analyses necessarily need to be included in the measurement model. If the variables in the measurement model are sufficient to model the relationship of a variable of interest and competence, the respective variable of interest does not need to be included in the measurement model for estimating plausible values. If, for example shoe size, weight, and clothing size are included in the measurement model for estimating plausible values for a certain competence, analyses on relationship estimates of the competence (using the respective plausible values) with body height are very likely to be unbiased. This is because body weight is well explained by the other three variables.

Many large-scale studies, such as the National Assessment of Educational Progress (NAEP) and the Programme for International Student Assessment (PISA), make use of this property. In these studies, a large number of context variables are present. In NAEP (Allen, Carlson, & Zelenak, 1999) and PISA (OECD, 2012), the context variables are included in the measurement model in form of orthogonal factors. In a factor analysis, as many orthogonal factors are extracted from the conditioning variables as required to explain at least 90% of the variance of the conditioning variables. The factor scores are then included in the measurement model for estimating plausible values. Although not all context variables are included in the measurement model, since the context variables are well explained by the factors, the respective plausible values may well be used to estimate most of the relationships of the competence scores and the context variables.

Once the desired plausible values are estimated, one may proceed with the analyses as described in section 5.2.

References

Adams, R. J., & Carstensen, C. H. (2002). Scaling outcomes. In R. J. Adams & M. Wu (Eds.). *PISA 2000 Technical Report* (p. 149-162). Paris: OECD.

Adams, R. J., Wu, M., & Carstensen, C. H. (2007). Application of multivariate Rasch Models in international large scale survey assessments. In M. Von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models - Extensions and applications* (p.271-280). New York: Springer.

Adams, R. J., Wilson, M. R., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1-24

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park: Sage.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716-722.

Allen, N.L., Carlson. J., & Zelenak, C.A. (1999). *The NAEP 1996 Technical Report* (NCES 99-452). Washington, DC: National Center for Education Statistics.

Andrich, D. (1985). A latent trait model for items with response dependencies: Implications for test construction and analysis. In: Embretson, S. E. (Ed.): *Test Design - Developments in Psychology and Psychometrics*. Orlando, San Diego, New York: Academic Press, Inc.

Aßmann, C., Carstensen, C. H., Gaasch, J.-C., & Pohl, S. (2012). *Estimation of plausible values using partially missing context variables – A data augmented MCMC approach*. Manuscript submitted for publication.

Aßmann, C., Steinhauer, H. W., Kiesel, H., Koch, S., Schönberger, B., Müller-Kuller, A., Rohwer, G., Rässler, S., & Blossfeld, H.-P. (2011). Sampling design of the National Educational Panel Study: Challenges and solutions. In H. P. Blossfeld, H. G. Roßbach & J. v. Maurice & (Eds.). *Education as a Lifelong Process: The German National Educational Panel Study (NEPS). Zeitschrift für Erziehungswissenschaft, Sonderheft 14* (p. 87-101). Wiesbaden: VS Verlag für Sozialwissenschaften.

Baumert, J. & Bos, W. (2000). *TIMSS/III – Dritte internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn*. Opladen: Leske + Budrich.

Behrendes, K., Weinert, S., Zimmermann, S., & Artelt, C. (2012). *Assessing language indicators across the life span within the German National Educational Panel Study (NEPS).* Manuscript submitted for publication.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F. M. & Novick, M. R. (Eds.). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Carstensen, C. H. (in press). Linking PISA competences over three cycles – results from Germany. In M. Prenzel, M. Kobarg, K. Schöps & S. Rönnebeck (Eds.). *Research outcomes of the PISA research conference 2009*. New York: Springer.

Carstensen, C. H., Knoll, S., Rost, J., & Prenzel, M. (2004). Technische Grundlagen. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, H.-G. Rolff, J. Rost, & U. Schiefele (Hrsg.). *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland - Ergebnisse des zweiten internationalen Vergleichs* (p. 371-387). Münster: Waxmann.

Clauser, B. E. & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items: An NCME instruction module. *Educational Measurement: Issues & Practice, 17*, 31-44.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analyses for the behavioral sciences* (3rd ed.). Hillsdale, NJ: Erlbaum.

Culbertson, M. (2011, April). Is it wrong? Handling missing responses in IRT. Paper presented at the Annual Meeting of the National Council of Measurement in Education, New Orleans, USA.

De Ayyala, R. J., Plake, B. S., Impara, J. C. (2001). The impact of omitted responses on the accuracy of ability estimation in Item Response Theory. *Journal of Educational Measurement, 38*, 213-234.

Elley, W. B. (1994). *The IEA study of reading: Achievement and instruction in thirty-two school systems*. Exeter: Pergamon.

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement, 45*, 225-245.

Frahm, S., Goy, M., Kowalski, K., Sixt, M., Strietholt, R., Blatt, I., Bos, W., & Kanders, M. (2011). Transition and development from lower secondary to upper secondary school. In H. P. Blossfeld, H. G. Roßbach & J. v. Maurice (Eds.). *Education as a Lifelong Process: The German National Educational Panel Study (NEPS). Zeitschrift für Erziehungswissenschaft, Sonderheft 14* (p. 217-232). Wiesbaden: VS Verlag für Sozialwissenschaften.

Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2012). *Framework for the assessment of reading competence within the NEPS. Dimensionality of the adult reading competence test.* Manuscript submitted for publication.

Glas, C. A. W., & Pimentel, J. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement, 68*, 907-922.

Gräfe, L. (2012). *How to deal with missing responses in competency tests? – A comparison of data- and model-based IRT approaches*. Unpublished master's thesis, Friedrich-Schiller-University Jena, Jena, Germany.

Haberkorn, K., Pohl, S., Carstensen, C. H., & Wiegand, E. (2012). *Incorporating different response formats in the IRT-scaling model for competence data.* Manuscript submitted for publication.

Hahn, I., Schöps, K., Rönnebeck, S., Martensen, M., Hansen, S., Saß, S., & Dalehefte, I. M. (2012). *Assessing science competency over the lifespan - A description of the NEPS science framework and the test development.* Manuscprit submitted for publication.

Holman, R. & Glas, C. A. W. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology, 58*, 1-17.

Kristen, C., Edele, A., Kalter, F., Kogan, I., Schulz, B., Stanat, P., & Will, G. (2011). The education of migrants and their children across the life course. In H. P. Blossfeld, H. G. Roßbach & J. v. Maurice & (Eds.). *Education as a Lifelong Process: The German National Educational Panel Study (NEPS). Zeitschrift für Erziehungswissenschaft, Sonderheft 14* (p. 121-138). Wiesbaden: VS Verlag für Sozialwissenschaften.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: J. Wiley & Sons.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39*, 247-264.

Ludlow, L. H. & O'leary, M. (1999). Scoring omitted and not-reached items: Practical data analysis implications. *Educational and Psychological Measurement, 59*, 615-630.

Macdonald, K. (2008/2011). Stata module to perform estimation with plausible values. Statistical Software Components, Boston College Department of Economics.

Maddala, G. S. (1977). *Econometrics*. New York: MC Graw-Hill, Inc.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49*, 359–81.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*(2), 177–196.

Mislevy, R. J. & Wu, P.-K. (1988). *Inferring examinee ability when some item responses are missing* (ERIC Document Reproduction Service No. ED 395 017). Princeton, NJ: Educational Testing Service.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Muthén, L. K. and Muthén, B. O. (1998-2010). *Mplus User's Guide. Sixth Edition*. Los Angeles, CA: Muthén & Muthén.

Neumann, I., Duchardt, C., Ehmke, T., Grüßing, M., Heinze, A., & Knopp, E. (2012). *Modeling and Assessing of Mathematical Competence over the Lifespan – Comparing Grade 9 Students and Adults.* Manuscript submitted for publication.

OECD (2005). *Pisa 2003 Technical Report*. Paris: OECD Publishing.

OECD (2009a). *PISA Data Analysis Manual: SAS, Second Edition*. OECD Publishing.

OECD (2009b). *PISA Data Analysis Manual: SPSS, Second Edition*. OECD Publishing.

OECD (2012). *PISA 2009 Technical Report*. OECD Publishing.

O'Muircheartaigh, C. & Moustaki, I. (1999). Symmetric pattern models: a latent variable approach to item non-response in attitude scales. *Journal of the Royal Statistical Society, Series A, 162*, 177-194.

Paulus, C. (2009). *Die "Bücheraufgabe" zur Bestimmung des kulturellen Kapitals bei Grundschülern.* Retrieved September 13, 2012, from Universität des Saarlandes: http://psydok.sulb.uni-saarland.de/volltexte/2009/2368/

Pohl, S. & Carstensen, C. (2012). *Scaling the competence tests in the National Educational Panel Study – Many questions, some answers, and further challenges*. Manuscript submitted for publication.

Pohl, S., Gräfe, L., & Rose, N. (2012). *Missing responses in competence tests – Explaining test taking strategies*. Manuscript in preparation.

Pohl, S., Gräfe, L., & Hardt, K. (2011). Ignorability und Modellierung von fehlenden Werten in Kompetenztests. Oral presentation given at the 10th Meeting of the Section Methods and Evaluation Research of the DGPs, Bamberg, Germany.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielsen & Lydiche (Expanded Edition, Chicago, University of Chicago Press, 1980).

Raudenbush, S.W., Bryk, A. S, & Congdon, R. (2004). *HLM 6 for Windows* [Computer software]. Skokie, IL: Scientific Software International, Inc.

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT)*. (ETS Research Report ETS RR-10-11), Princeton, NJ: Educational Testing Service.

Rosenbaum, P. R. (1988). A note on item bundles. *Psychometrika, 53*, 349-360.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

Senkbeil, M., Ihme, J. M., & Wittwer, J. (2012). *The test of technological and information literacy (TILT) in the National Educational Panel Study: Development, empirical testing, and preliminary evidence for validity.* Manuscript submitted for publication.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237-247.

Stanat, P. & Segeritz, M. (2009). Migrationsbezogene Indikatoren für eine Bildungsberichterstattung. In R. Tippelt (Hrsg.). *Steuerung durch Indikatoren? Methodologische und theoretische Reflexionen zur deutschen und internationalen Bildungsberichterstattung* (S. 141–156). Opladen: Barbara Budrich.

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are Plausible Values and why are they useful? In M. von Davier and D. Hastedt (Eds.), *IERI monograph series: Issues and methodologies in large scale assessments (vol. 2).* IEA-ETS Research Institute.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admission Test as an example. *Applied Measurement in Education, 8*, 157-186.

Wainer, H. & Kiely (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.

Wainer, H. & Lukhele, R. (1997). How reliable are TOEFL scores? *Educational and Psychological Measurement, 57*, 749-766.

Wainer, H. & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*, 22-29.

Wainer, H. & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37*, 203-220.

Walter, O. (2005). *Kompetenzmessung in den PISA-Studien. Simulation zur Schätzung von Verteilungsparamtern und Reliabilitäten*. Lengerich: Pabst.

Wang, W.-C. & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29*, 126-149.

Warm, T. A. (1989). Weighted likelihood estimation of ability in Item Response Theory. *Psychometrika, 54*, 427-450.

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T. & Carstensen C. H. (2011). Development of competencies across the life span. In H. P. Blossfeld, H. G. Roßbach & J. v. Maurice & (Eds.). *Education as a Lifelong Process: The German National Educational Panel Study (NEPS). Zeitschrift für Erziehungswissenschaft, Sonderheft 14* (p. 67-86) . Wiesbaden: VS Verlag für Sozialwissenschaften.

Wilson, M. (2003). On choosing a model for measuring. *Methods of Psychological Research, 8*, 1-22.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

Wilson, M. & deBoeck, P. (Eds) (2004). *Explatanatory item response models*. New York: Springer.

Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement.* Chicago: MESA Press.

Wu, M. L. (1997). *The development and application of a fit test for use with marginal maximum likelihood estimation and generalized item response models*. Unpublished Masters Dissertation. University of Melbourne.

Wu, M., Adams, R. J., Wilson, M. & Haldane, S. (2007) *Conquest 2.0*. Camberwell: ACER Press.

Yen, W. (1993). Scaling performance assessment: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.

Zhang, O., Shen, L., Cannady, M. (2010). Polytomous IRT or testlet model: An evaluation of scoring models in small testlet size situations. Paper Presented at Annual Meeting of the 15th International Objective Measurement Workshop. Boulder, Colorado.

Ziesky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337-348): Hillsdale, NJ: Erlbaum.

Appendix

Appendix A: Response formats in the competence test

a) Simple multiple choice

There are countries in the European Union that are smaller than Luxemburg. How many?

Please tick the right answer! Please tick just one answer!

O	Only one country is smaller than Luxemburg.
O	Two countries in the European Union are smaller than Luxemburg.
O	Four countries are smaller than Luxemburg.
O	Five countries are smaller than Luxemburg.

b) Complex multiple choice

What do you get to know in the text about Luxemburg?

Decide for each row whether the statement is right or wrong!

	right	wrong
a) In the text, there is information about the size of the country.	O	O
b) The text reports on the history of the country.	O	O
c) In the text, they inform about the currency in Luxemburg.	O	O

c) Matching item

Sort the headings to the corresponding passages in the text!

Passages		Headings	
1.	☐	A	Luxemburg and the EU
2.	☐	B	Location and size of Luxemburg
3.	☐	C	Luxemburg as the financial center
4.	☐	D	Government and inhabitants of Luxemburg
		E	The cuisine of Luxemburg

d) Short-constructed response

Calculate the area of the square above!

Area = ☐ cm^2

Appendix B: ConQuest-Syntax and output for investigating the effect of gender on ability

Syntax

```
Title effect of gender on ability;
data filename.dat;
format responses 1-20 position 22 gender 24; /* insert number of columns with data*/
labels << filename_with_labels.txt;

codes 0,1;
score (0,1) (0,1)                                !items (1-20);

set constraint = cases;
regression gender;
model item + item*step + position;
estimate;
show !estimates=latent >> filename.shw;
```

Output

```
Regression Variable
CONSTANT        0.000*
Gender          -0.175 (0.032)
-----------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
===============================================
COVARIANCE/CORRELATION MATRIX
Dimension
Dimension 1
----------------------------------
Variance        1.360
----------------------------------
```

Appendix C: ConQuest-Syntax for estimating plausible values

```
Title Estimating Plausible Values;
data filename.dat;
format responses 1-20 position 22 gender 24 school 26 age 28; /* insert number of columns with data*/
labels << filename_with_labels.txt;

codes 0,1;
score (0,1) (0,1)                    !items (1-20);

set constraint = cases;
regression gender school age;
model item + position;
estimate;

show !estimates=latent >> filename.shw;
show cases !estimates=latent >> filename.pv;
```