Journal for Educational Research Online
Journal für Bildungsforschung Online
Volume 5 (2013), No. 2, 189–216

Steffi Pohl & Claus H. Carstensen

Scaling of competence tests in the National Educational Panel Study – Many questions, some answers, and further challenges

Abstract

Competence measurement in (longitudinal) large-scale assessments, such as in the National Educational Panel Study (NEPS), imposes specific demands on the scaling of the competence data. These challenges include scaling issues such as the question on how to deal with different response formats as well as with missing values in the scaling model. They also include design aspects, as for example, the incorporation of adaptive testing and the inclusion of students with special educational needs in the assessment. Especially in longitudinal designs the question of linking of competence scores across different cohorts and measurement occasions arises. With this article we aim at pointing out some of the challenges one has to meet when scaling competence data of (longitudinal) large-scale assessments, at giving an overview of research we have conducted within the NEPS to find solutions to these questions, and at pointing out some directions for future research. While for most of the topics we give an overview of the research that has been conducted in NEPS, we more thoroughly describe the research we have conducted on investigating the assumptions necessary for linking of different cohorts in NEPS. We specifically target the question whether the same competence may be measured coherently across the whole lifespan. The results show that for linking the Grade 9 reading test to the adult reading test, measurement invariance does not hold for all items. The implementation of the research results into the scaling of NEPS competence data is described and the applicability of the research results to other large-scale studies is discussed.

Keywords

Competence tests; Item Response Theory; Scaling; Linking; Lifespan

Dr. Steffi Pohl (corresponding author), National Educational Panel Study (NEPS), University of Bamberg, Wilhelmsplatz 3, 96045 Bamberg, Germany
e-mail: steffi.pohl@uni-bamberg.de

Prof. Dr. Claus H. Carstensen, Chair of Psychology with a Focus on the Methods of Empirical Educational Research, University of Bamberg, Wilhelmsplatz 3, 96045 Bamberg, Germany
e-mail: claus.carstensen@uni-bamberg.de

Skalierung der Kompetenztests im Nationalen Bildungspanel – Viele Fragen, einige Antworten und weitere Herausforderungen

Zusammenfassung

Kompetenzmessung in (längsschnittlichen) groß angelegten Erhebungen wie dem Nationalen Bildungspanel (NEPS) stellen spezielle Anforderungen an die Skalierung der Kompetenztestdaten. Die Herausforderungen beinhalten Skalierungsfragen, wie die Frage nach der Berücksichtigung verschiedener Antwortformate und die Behandlung von fehlenden Werten im Skalierungsmodell. Sie beinhalten aber auch Designaspekte, wie zum Beispiel die Umsetzung von adaptivem Testen und die Inklusion von Schülern mit sonderpädagogischem Förderbedarf in die Kompetenzerhebung. Besonders in längsschnittlichen Designs stellt sich auch die Frage nach der Verlinkung von Kompetenzwerten über verschiedene Kohorten und Messzeitpunkte. Mit diesem Artikel wollen wir einige der Herausforderungen darlegen, denen man sich bei der Skalierung der Kompetenztestdaten stellen muss, einen Überblick über Forschung geben, die wir im Rahmen des NEPS durchgeführt haben, um Antworten auf diese Fragen zu finden, sowie Themen für weitere Forschung aufzeigen. Während wir für den Großteil der Themen einen Überblick über die bisher durchgeführte Forschung geben, beschreiben wir die Forschung zur Untersuchung der Annahmen für eine Verlinkung von Kohorten im NEPS ausführlicher. Im Speziellen untersuchen wir ob dieselbe Kompetenz über die gesamte Lebensspanne kohärent gemessen werden kann. Die Ergebnisse zeigen, dass für die Verlinkung des Lesetests der neunten Klasse mit dem Test der Erwachsenen Messinvarianz nicht für alle Items gilt. Die Implementation der Forschungsergebnisse in die Skalierungspraxis der Kompetenztestdaten im NEPS wird erläutert und die Anwendbarkeit der Forschungsergebnisse für andere groß angelegte Studien wird diskutiert.

Schlagworte

Kompetenztests; Item-Response-Theorie; Skalierung; Verlinkung; Lebensspanne

1. Introduction

A particular strength of the National Educational Panel Study (NEPS) is that it collects detailed competence data in a large, longitudinal, multicohort sequence survey design as well as detailed data on conditions for and consequences of individual educational careers. Thus, a wide range of research questions regarding the development of competencies as well as the interaction between competence development and context factors with respect to individual educational careers may be investigated (see, e.g., Blossfeld, von Maurice, & Schneider, 2011). As presented by Artelt, Weinert, and Carstensen (2013, this issue), the framework for assessing

competencies in the NEPS employs a number of different domains. These include, among others, reading comprehension, mathematical competence, and scientific literacy, as well as information and communication technologies (ICT) literacy. In order to analyze competencies and their relations to other variables without disattenuation by measurement error, a latent variable modeling approach has been applied. With the scaling of competence tests we aim to provide reliable competence scores that are purified from measurement error and that allow researchers to investigate latent relationships of competence scores with explaining variables. The assessment of competencies is based on test instruments that require participants to respond to single tasks in various tests. Whereas Gehrer, Zimmermann, Artelt, and Weinert (2013, this issue), Neumann et al. (2013, this issue), Hahn et al. (2013, this issue), and Senkbeil, Ihme, and Wittwer (2013, this issue) elaborate how the respective domain-specific competence tests are developed, this article discusses the challenges to be met when analyzing competence data. Note that different scaling approaches are used for different competence tests. Whereas established tests in NEPS (such as those measuring reading speed, basic cognitive skills, listening comprehension at word level) are usually scaled according to the test manual via classical test theory, newly developed tests (such as those for reading comprehension, mathematical competence, scientific literacy, and ICT literacy) in NEPS are usually scaled on the basis of Item Response Theory (IRT). This article deals with the challenges arising in IRT scaling of competence tests in (longitudinal) large-scale assessments and research conducted to find solutions for these questions.

IRT was chosen as scaling framework for the newly developed tests because it allows for an estimation of item parameters independent of the sample of persons and for an estimation of ability independent of the sample of items. With IRT it is possible to scale the ability of persons in different waves on the same scale, even when different tests were used at each measurement occasion. Sophisticated measurement models have been developed within the IRT framework and are frequently used in large-scale assessments. In fact, the state of the art in analyzing test data from large-scale assessment programs is to combine appropriate measurement models for item responses on the one hand and sophisticated statistical models for the structural part of the model – such as generalized linear modeling, structural equation modeling, or multilevel modeling – on the other hand. Hence, complex models that explicitly include the measurement model of competencies may be specified for answering research questions. The strength of such an approach is that various sources of error (e.g., measurement error, imputation error) are simultaneously accounted for in the model.

For scaling the NEPS competence data (see Pohl & Carstensen, 2012 for a detailed description of the scaling of competence data in the NEPS), the Rasch model (Rasch, 1960/1980) was chosen for dichotomous items and the Partial Credit Model (PCM, Masters, 1982) was chosen for polytomous items. All models were estimated in ConQuest (Wu, Adams, Wilson, & Haldane, 2007). The implementation of the tests as well as the scaling of the tests imposes several demands on

psychometric models to capture the properties of the data and provide unbiased and precise parameter estimates. One is to construct a test that is appropriate in its difficulty for the specific target group. Challenges may also arise from the test construction principles (see Gehrer et al., 2013, this issue; Hahn et al., 2013, this issue; Neumann et al., 2013, this issue; Senkbeil et al., 2013, this issue). The NEPS tests employ different response formats, common stimuli for some items induce local dependencies among these items, and different kinds of missing responses occur. These different aspects need to be accounted for in the scaling model. Furthermore, the NEPS is a longitudinal study with a multicohort sequence design aimed at investigating the change of competencies over time. In order to investigate change, competencies need to be measured coherently across the whole lifespan – from kindergarten to adult age – and competence scores need to be statistically linked onto the same scale.

Accounting for the different test construction principles, the complex study design, and the broad demands placed on the released data, the data of large-scale assessments – especially those with a longitudinal design – impose specific demands on the scaling model that need to be met. In this paper we point out some of the questions that need to be addressed when working with competence data in longitudinal large-scale studies such as the NEPS. The challenges discussed are those (a) that arise in large-scale assessments, especially those with a longitudinal design, (b) for which no satisfactory solution exist in the literature as yet, and (c) that have been approached in the NEPS. These include specific aspects of the scaling model, that is, (1) dealing with different response formats and (2) the treatment of missing responses as well as further aspects of scaling, that is, (3) adaptive testing, (4) testing students with special educational needs, and (5) linking across cohorts. While giving an overview of research that we have conducted on the first four topics, we present in more detail research on the comparability of test scores across age cohorts. In the following sections we first present an overview of research on the first four topics. For each of the four topics we give a short overview of the state of the art and current practice in large-scale assessments. We then shortly present research conducted in NEPS to further develop these approaches and also describe how these research results are subsequently implemented in NEPS. As an example, we introduce more specifically the research design and the research results on the linking of test scores across cohorts. Although the research presented here has been conducted on NEPS data, the results may also be valuable for other large-scale studies as well as for researchers working with data from large-scale studies. This point is being elaborated as part of the discussion.

2. Overview of research on the design and scaling of competence data

2.1 Incorporating different response formats

To assess competencies in large-scale studies such as the Programme for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS), and the National Assessment of Educational Progress (NAEP), the responses to a number of tasks are recorded. Usually, different response formats are used in these tests (e.g., Allen, Donoghue, & Schoeps, 2001; OECD, 2012; Olson, Martin, & Mullis, 2008). In NEPS, assessing competencies in each domain usually takes about 28–30 minutes and between 20 and 35 items. Most of the items in NEPS (see Gehrer et al., 2013, this issue; Hahn et al., 2013, this issue; Neumann et al., 2013, this issue; Senkbeil et al., 2013, this issue) have a simple multiple-choice (MC) format with four response options (see Figure 1a). Further response formats are complex multiple-choice (CMC) items, matching items, and short constructed responses. Complex multiple-choice items present a common stimulus followed by a number of MC questions with two response options each (see Figure 1b). Matching (MA) items consist of a common stimulus followed by a number of statements, which require assigning a list of response options to these statements (see Figure 1c). These questions are typically used for matching titles to different paragraphs of a text in reading assessments. Short constructed responses are used in mathematics tests only. They usually present a mathematical problem that requires a numerical answer (see Figure 1d). The question is how these different response formats may adequately be incorporated in the scaling model. In this section we first focus on local item dependence (LID) introduced by specific response formats and how aggregation of locally dependent items to polytomous super items may account for LID. We then discuss how different response formats can be weighted in the measurement model when estimating the competence score and how different item formats were dealt with in scaling the data of the NEPS.

Whereas simple MC items and short constructed responses are treated as dichotomous variables, complex MC items and matching items consist of a number of items that share a common stimulus. These item bundles pose problems for the assumption of local stochastic independence (see, e.g., Yen, 1993). This is especially prevalent in the matching tasks. As response options are to be used only once in these tasks, a response to one subitem[1] heavily depends on the responses to previous subitems. 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

1 We refer to subitems as the data of the single statements of a complex MC or matching task that form an item bundle.

Figure 1: Response formats in the competence tests – examples (note that these are just illustrative examples and not items that are used in the actual tests): a) simple multiple-choice, b) complex multiple-choice, c) matching task, and d) short constructed response

a) Simple multiple-choice

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

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

☐ Only one country is smaller than Luxembourg.

☐ Two countries in the European Union are smaller than Luxembourg.

☐ Four countries are smaller than Luxembourg.

☐ Five countries are smaller than Luxembourg.

b) Complex multiple-choice

What do you get to learn from the text about Luxembourg? Decide for each row whether the statement is right or wrong!

		right	wrong
a)	The text gives information about the size of the country.	☐	☐
b)	The text reports on the history of the country.	☐	☐
c)	In the text they talk about the currency in Luxembourg.	☐	☐

c) Matching task

Match the headings to the respective passages in the text!

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

d) Short constructed responses

Calculate the area of the square above!

Area = ☐ cm^2

& Lukhele, 1997; Wainer & Thissen, 1996; Wainer & Wang, 2000; Yen, 1993). Researchers (e.g., Andrich, 1985; Zhang, Shen, & Cannady, 2010) recommended the use of super items that are aggregates of the subitems analyzed via, for example, the PCM (Masters, 1982) for tests with a large number of testlets and a small testlet size. In line with other large-scale studies (e.g., PISA; Adams & Wu, 2002; OECD, 2009) and previous research (e.g., Andrich, 1985; Wilson, 1988; Wilson & Adams, 1995; Zhang et al., 2010), in the NEPS we accounted for item dependence of item bundles by aggregating the subitems of CMC and MA items to an ordered polytomous score (super item) and analyzed the data via the PCM (Pohl & Carstensen, 2012). The values of the aggregated scores give the number of correctly answered subitems. They range from zero (no correct answer) to the number of subitems (all subitems answered correctly). Analyzing responses from different response formats in one-parameter (1PL) models has interesting implications on the assumed discrimination of items with different formats. Since we do not estimate discrimination parameters in the PCM (as in a two-parameter (2PL) model, Birnbaum, 1968; Muraki, 1992), the weight of the items is modeled solely by the scoring of the responses. The higher an item is scored, the higher its discrimination. In terms of scaling the data, the question arises as to how the response categories of polytomous CMC and MA variables should be scored in a scaling model and how the different item formats should be weighted? Whereas in our tests simple MC items consist of four alternatives with one correct answer, complex MC items consist of a number of subitems with two response alternatives each. Furthermore, as the number of subitems in complex MC formats and matching tasks varies (from two to seven), the maximum number of score points of the super items also varies. Should CMC and MA items with more subitems have more impact on the overall competence score than items with fewer subtasks? Or should all items receive the same maximum score, regardless of the response format and number of subtasks, and, thus, have the same impact on the overall competence score? An appropriate approach to scoring ought to reflect the amount of information obtained from the responses.

In PISA, complex MC items were aggregated to dichotomous or polytomous super items (e.g., OECD, 2009, 2012). In the analyses, the different values of the polytomous variables were collapsed into two to four categories. Which of the values were collapsed into one category was decided on the basis of theoretical and statistical arguments. Each category was then scored as one score point, thus resulting in polytomous variables with a maximum score varying between 1 (for two categories) to 3 (for four categories). However, also other scoring rules (see, e.g., Ben-Simon, Budescu, & Nevo, 1997) exist (see Table 1). First, the super items may be recoded to a dichotomous variable, where a score of 1 is assigned when all subitems are answered correctly and a score of 0 otherwise (all-or-nothing-scoring). This would imply that all items and response formats have the same weight (discrimination). Another scoring rule may make the assumption that each subitem of a CMC or MA item has the same weight as a simple MC item (number-correct scoring with one point per correct subtask). Thus, each correct answer on the

subitems of a polytomous super item would result in one score point. The maximum score of a polytomous item would then be equal to the number of subitems forming the super item. Another quite plausible assumption is that the subitems of complex items do not have the same weight as simple MC items but a reduced weight, for example, 0.5 points per correct answer (number-correct scoring with half points per correct subtask). This would account for the reduced number of response options (two instead of four in simple MC items) in CMC items as well as for the dependence of subitems in matching tasks. A scoring of one or half points per subtask implies that super items with many subitems receive a higher weight than super items with fewer subitems. Up to now, there has been only little research on *weighting different types of item formats*. The only research regarding weighting so far has focused solely on *weighting items*. This research was mostly conducted within the framework of classical test theory (e.g., Ben-Simon et al., 1997; Wongwiwatthananukit, Bennett, & Popovich, 2000) and hardly any research has dealt with this topic at all within the framework of IRT (Lukhele & Sireci, 1995; Si, 2002; Sykes & Hou, 2003).

Table 1: Example of different scoring methods of a CMC item with 5 categories

	Scoring rule		
Categories of a CMC item with five subtasks	All-or-nothing scoring	Number-correct scoring with half points per correct subtask	Number-correct scoring with one point per correct subtask
0	0	0	0
1	0	0.5	1
2	0	1	2
3	0	1.5	3
4	1	2	4

Haberkorn, Pohl, Carstensen, and Wiegand (2013) investigated the fit of the different scoring rules to the competence data in the NEPS. They found that a higher discrimination of the test results when a number-correct scoring rule, which differentiates between categories, is applied. Collapsing the categories into a dichotomous variable (all-or-nothing-scoring) leads to loss of information. Haberkorn et al. (2013) furthermore investigated which of the remaining scoring rules best describe the competence data in the NEPS. The resulting weighted mean square error (WMNSQ, Wright & Masters, 1982) using one point or half a point per correct subtask for reading competence (see Gehrer et al., 2013, this issue for the description of the framework) assessed in adults and for scientific literacy (see Hahn et al., 2013, this issue for the description of the framework) assessed in Grade 9 students show overall better fit values (mean WMNSQ near 1 and small standard deviation) for half-point-per-correct-subtask than for one-point-per-subtask scoring. This indicates that a scoring of 0.5 for each category in complex multiple-choice format items better models their impact on competence scores in our tests than a scoring

of 1 for each of these categories. These results were also confirmed by 2PL-analyses of the items. In the respective analyses estimated discrimination parameters closely resembled the discrimination assumed by half-point-per-correct-subtask scoring. The results were consistent for different competence domains (reading competence and scientific literacy, as well as ICT) and different age cohorts (students in Grade 5, Grade 9, as well as adults). Hence, in the NEPS half-point-per-correct-subtask scoring was implemented in the scaling model. The results show that the different response formats need to be incorporated differently in the scaling model. For response formats used in the NEPS, a half-point-per-subitem scoring of complex MC items described the data best. In other large-scale studies with other response formats, different scoring options could theoretically be developed and empirically be tested.

2.2 Dealing with missing responses

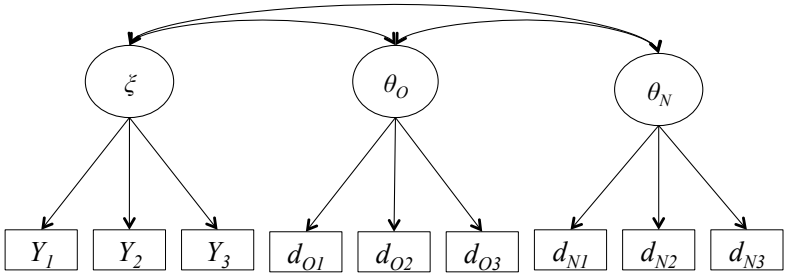
Usually, missing responses occur in competence data. Responses to competence items may be missing due to different reasons. These are (1) items that were not administered (due to the testing design), (2) invalid responses (e.g., more than one response to a simple MC item), (3) omitted items, and (4) items that were not reached due to time limits (i.e., omitted responses after the last given response). The amount of missing responses in large-scale assessments is usually not negligible. In PISA 2006, for example, across all countries and all three domains (mathematics, reading, and science) on average about 10% (in Germany 8.37%) of items were omitted and 4% (in Germany 1.15%) of the items were not reached (OECD, 2009, pp. 219–220). For mathematics and science in TIMSS 2003, on average 3.73% of the items were not reached in Grade 8 and 5.96% in Grade 4 (Mullis, Martin, & Diaconu, 2004, p. 249).

The ignorability of the missing responses depends on the reason for responses to be missing. Whereas in most test designs missing responses due to not-administered items are missing by design and, therefore, missing completely at random or missing at random, omitted and not-reached items are usually nonignorable and often depend on the difficulty of the item and the ability of the person (Mislevy & Wu, 1988). If not treated correctly, nonignorable missing responses may lead to biased parameter estimates (Mislevy & Wu, 1996) and, thus, wrong conclusions about competence levels of persons as well as about relationships of competencies with other variables.

Different approaches for dealing with missing responses in competence tests exist. Missing responses may, for example, be treated as not administered, as incorrect, or as fractionally correct (see, e.g., Culbertson, 2011; Lord, 1974). While in NAEP omitted items are scored as fractional correct, items that were not reached are treated as not administered and are, thus, ignored in the parameter estimation (Allen, Carson, Johnson, & Mislevy, 2001). In some large-scale studies, such as PISA, TIMSS, and the Progress in International Reading Literacy Study (PIRLS),

there is a two-stage procedure for treating missing responses. For the estimation of the item parameters missing responses are treated as not administered (ignored in the parameter estimation). The estimated item parameters are then used as fixed parameters for the estimation of person parameters, where missing responses are scored as incorrect (e.g., Macaskill, Adams, & Wu, 1998; Martin, Mullis, & Kennedy, 2007; OECD, 2009). Whereas all of these approaches rely on the assumption that the missing reponses are ignorable, model-based approaches for nonignorable missing data (Holman & Glas, 2005; Glas & Pimentel, 2008; Moustaki & Knott, 2000; O'Muircheartaigh & Moustaki, 1999) have recently been developed. In these approaches a latent missing propensity is modeled and included as a conditioning variable in the measurement model. A model that includes both a latent missing propensity for omitted and a latent missing propensity for not-reached items (see Pohl, Gräfe, & Rose, in press) is depicted in Figure 2. Note that, while Y_i represents the responses on the competence items, d_{Oi} and d_{Ni} represent missing indicators indicating whether the response to an item i is omitted or not reached, respectively. ξ represents the latent competence score and θ_O and θ_N the missing propensity due to omission and speed, respectivley.

Figure 2: Model-based approach of treating different kinds of nonignorable missing responses (Pohl, Gräfe, & Rose, in press). Y_i indicates the responses to the competence items i, d_{Oi} and d_{Ni} indicate the missing response indicator for omission and not reaching the item, respectively, for each item i, ξ indicates the latent competence, θ_O, and θ_N the latent missing propensity due to omission and speediness, respectively.



As Lord (1974) as well as Mislevy and Wu (1988, 1996) analytically derived, scoring missing responses as incorrect violates the model assumptions of IRT models. It induces a deterministic term and results in local item dependence. This has been corroborated in simulation studies and empirical analyses (e.g., Finch, 2008; Rose, von Davier, & Xu, 2010), which have shown that treating missing responses as incorrect results in biased estimates of item and person parameters. Ignoring omitted and not-reached items or using the model-based approaches leads to unbiased item and person parameter estimates. Model-based approaches do result in a higher reliability and allow us to investigate the ignorability of the missings. They are,

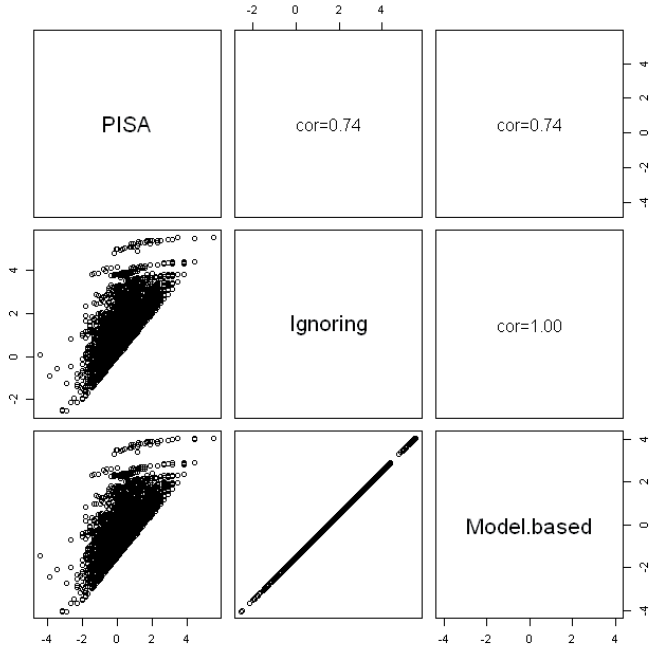
however, more complex and require model assumptions (such as unidimensionality of the missing responses).

Most of the previous results are based on analytical derivations or simulation studies. Not much research has investigated the suitability of the different approaches for empirical data. Thus, Pohl, Gräfe, and Rose (in press) investigated the performance of different approaches for dealing with missing values on the parameter estimates for reading (see Gehrer et al., 2013, this issue) and mathematical competence (see Neumann et al., 2013, this issue) measured in Grade 5 in the NEPS. The respective reading test in NEPS consists of 33 items (with a total of 59 subitems, see Pohl, Haberkorn, Hardt, & Wiegand, 2012), whereas the mathematics test consists of 25 items (28 items when considered on the subitem level, see Duchhardt & Gerdes, 2012). In both tests, the number of nonvalid responses is very low (M < 0.15%). The subjects omit on average about 5.3% (SD = 8.7%) of the items. The omission rate does not differ between the mathematics and the reading test. However, while reading shows a large amount of not-reached items (M = 13.7%, SD = 16.9%), hardly any items were not reached in the mathematics test (M = 1.5%, SD = 7.6%). Pohl et al. (in press) tested different approaches to dealing with the different types of missing responses. Among others, they compared the approach of PISA (ignoring missing responses for the estimation of item parameters and scoring them as incorrect for the estimation of person parameters), the approach of ignoring missing responses in the estimation of the parameters, and the model-based approach for nonignorable missing data for both not-reached and omitted items.

The estimated person parameter estimates (Warms Maximum Likelihood Estimates; WLE; Warm, 1989) for reading competence are presented in Figure 3. Although there are considerable correlations between the latent missing propensities for omitted and for not-reached items with ability (r = -.175 for omitted and r = .200 for not-reached items), indicating nonignorability, the person parameter estimates hardly differ between ignoring missing responses and using a model-based approach (r = 1). In empirical data, the approach of ignoring missing responses seems to be robust to violations of nonignorability. The person parameter estimates when ignoring missing responses differ considerably from those using the approach of PISA. Since the reading test shows a larger amount of missing responses than the mathematics test, the impact of the missing approach is more viable for the parameter estimates of reading competence than for mathematical competence. The person parameter estimates of the two approaches correlate with each other to r = .743 for reading and r = .954 for mathematical competence, indicating that the two approaches result in different ability estimates. Considerable differences in the person parameter estimates occur for those persons that have a large number of missing responses. The ability of these persons is, thus, heavily underestimated using the PISA approach as compared to ignoring the missing responses. The results found in empirical analyses using NEPS data are in line with results found in complete case simulation studies on the same data (Pohl et al., in press), thus supporting the superiority of the approach of ignoring missing responses as

well as model-based approaches. Since the results show that the approach of ignoring missing responses is robust to violations of ignorability in these applications, it was decided to ignore missing responses in the scaling model of the NEPS competence data. In large-scale studies measuring competencies, model-based approaches can be used to investigate the amount of nonignorability and – if parameter estimates differ only slightly between model-based approaches and ignoring missing responses – to justify ignoring missing responses in the scaling model.

Figure 3: WLE person parameter estimates for Grade 5 students on reading competence comparing treating missing responses as in PISA, ignoring missing responses, and using a model-based approach



2.3 Adaptive Testing

Large-scale assessments aim at precisely measuring the competencies of all persons. This is often a difficult endeavor because the persons in these studies usually show a wide range of ability. Furthermore, in order to reduce the burden on test takers, testing time is limited. However, it is quite difficult to measure the competencies of persons (a) across a wide ability range, (b) within a short time, (c) as accurately as possible, while (d) still trying to avoid a loss of motivation and, therefore, panel dropouts. These different requirements may be met by using adaptive testing, where the difficulty of items presented to a person is matched to the specif-

ic ability of that particular person. In NAEP (Xu, Sikali, Oranje, & Kulick, 2011) as well as in PISA (Pearson, 2011), conversion to adaptive testing is already planned.

However, sometimes organizational constraints do not allow for classical adaptive testing. At this stage, in NEPS, most of the tests are implemented in group test settings using paper-and-pencil mode and the number of items available for each competence domain is limited. Thus, computer-adaptive testing (CAT, e.g., Lord, 1971b; McKinley & Reckase, 1980; van der Linden & Glas, 2000, 2010; Wainer, 2000) or classical multi-stage testing (MST, Angoff & Huddleston, 1958; Cronbach & Gleser, 1965; Linn, Rock, & Cleary, 1969; Lord, 1971a) may not yet be applied. As a consequence, drawing on the idea of Beard (2008) an alternative adaptive testing design (longitudinal MST), that may be implemented as a paper-and-pencil test in group settings and does not need a large pool of calibrated items, has been developed for the NEPS by Pohl (in press). In longitudinal MST, competence scores from previous waves are used to allocate subjects to different test forms (which differ in their difficulty level) in later waves. This design makes use of information available in longitudinal studies. Pohl (in press) performed a simulation study investigating the bias and efficiency not only of the measurement of competence but also of competence development using longitudinal MST as opposed to conventional testing (one test form for all subjects). She found that longitudinal MST does indeed increase the measurement precision, especially for subjects at the lower and upper end of the ability distribution. Regarding the measurement precision of the change of competence scores across waves and the motivation of the subjects, the competence scores from the previous wave should correlate to at least $r = .7$ with the competence score in the later wave. If the correlation is smaller than $r = .7$, the number of misallocations of test forms to persons becomes rather large and the measurement precision of change scores is smaller for adaptive testing than for conventional testing. Misallocations should especially be avoided because they may result in a loss of motivation in test takers. Since competence measures are usually very stable across time (e.g., Prenzel et al., 2006; Rock, Pollack, & Quinn, 1995), pretest measures may well be used for routing to test forms. Longitudinal MST is currently implemented for the second measurement wave of the same competence in the NEPS. For the second measurement of reading and mathematics competence in Gade 7 and Grade 9, two test forms – an easy and a hard one – are administered to students on the basis of their competence scores 2 years ago.

2.4 Testing students with special educational needs

Many large-scale studies, such as NAEP, try to include students with special educational needs in the assessment of competencies (e.g., Lutkus, Mazzeo, Zhang, & Jerry, 2004). NEPS also aims at including students with special educational needs in its survey (Heydrich, Weinert, Nusser, Artelt, & Carstensen, 2013, this issue). As a specific group, the NEPS focuses on students with special educational needs in the area of learning (SEN-L). In order to be able to thoroughly investigate this

group, it is oversampled in the survey (Aßmann et al., 2011). The challenge is to provide students with SEN-L with tests and test conditions that will ensure a reliable and valid assessment of competence scores. A challenge is to obtain comparable competence measures for special education and general education students. Various studies (e.g., Koretz & Barton, 2003) show the difficulty of assessing the competence level of students with special educational needs (SEN) with the same test instrument as administered to general education students. Therefore, students with SEN are often tested using accommodated tests and testing conditions (e.g., Lutkus et al., 2004). Test accommodations for students with SEN-L usually include a reduction of item difficulty (out-of-level testing), frequent breaks, and extended testing time (e.g., Koretz & Barton, 2003). The provision of testing accommodations for individuals with disabilities is a highly controversial issue in the assessment literature (Pitoniak & Royer, 2001; Sireci, Scarpati, & Li, 2005). Objections to test accommodations regard the comparability of test results between students with SEN and general education students and, thus, test fairness (Abedi et al., 2011; Bielinski, Thurlow, Ysseldyke, Freidebach, & Freidebach, 2001; Cormier, Altman, Shyyan, & Thurlow, 2010). Test accommodations do not necessarily need to result in competence scores that are comparable to those gained from a standard test (see, e.g., Pitoniak & Royer, 2001). Reducing test length or test difficulty may alter the construct measured (e.g., Lovett, 2010). Up to now, research on the comparability of test scores has revealed ambiguous results (e.g., Bolt & Ysseldyke, 2008; Finch, Barton, & Meyer, 2009; Koretz, 1997; Lutkus et al., 2004). Lovett (2010) points out limitations of studies on the comparability of test results; these are (a) small sample sizes, (b) heterogeneous student groups and test accommodations, as well as (c) confounding of test accommodation and student group.

NEPS conducts feasibility studies in order to evaluate which tests and test settings serve to assess students with SEN-L reliably and validly as part of the study (Heydrich et al., 2013, this issue). Südkamp, Pohl, and Weinert (2013) investigated the comparability of reading competence scores estimated for general education and SEN-L students in the NEPS. They drew on a large, representative sample of students with SEN-L. Next to the standard reading test, two test accommodations featuring (a) a reduction in test length and (b) a reduction in test difficulty were evaluated in a control group of students in the lowest academic track (to evaluate the test accommodations themselves) as well as to students with SEN-L (to evaluate the appropriateness of the test versions for this specific target group). The results favor the implementation of a reduced test length because they led to a lower amount of missing responses as compared to the standard test for students with SEN-L. However, all versions of the reading test showed low item fit to the respective IRT model for students with SEN-L (14% to 38% of the items in the respective test versions showed unsatisfying point biserial correlations) and a considerable amount of measurement variance (differential item functioning, DIF) between students with SEN-L and general education students. The tests seemed to measure a different construct for both samples. As a consequence, an appropriate link between the test data of general eduction students and accommodated test

data of students with SEN-L that will enable a comparison of both samples is not possible. Currently, research is being conducted (Pohl, Hardt, Südkamp, Nusser, & Weinert, 2013) in a bid to identify students with SEN-L for whom the standard or the accommodated reading tests provide reliable and comparable reading competence scores. If this attempt succeeds, it will be possible to publish reliable competence scores for some of the SEN-L students in the Scientific Use File and enable research to investigate competence acquisition and development for this particular group of students.

3. Empirical study on linking of cohorts

In longitudinal large-scale studies, such as the NEPS, a particular aim is to investigate the development of competencies over the life course. In NEPS, the competence domains are repeatedly assessed after 2 to 6 years, allowing researchers to closely follow the development of competencies. In the multi-cohort-sequence design (see Artelt et al., 2013, this issue) different starting cohorts may also be compared. Linking of different cohorts includes the linking between cohorts at the same measurement occasion, for example, adults and Grade 9 students in the first wave, and linking of different cohorts at the same age, for example, linking of the results in Grade 9 between Starting Cohort 3 (starting in Grade 5) and Starting Cohort 4 (starting in Grade 9). If we assume that, at least within certain stages, the construct measured in a certain domain stays the same over time and only the extent to which participants show proficiency in that particular competence changes, a series of assessments might be constructed, which in operational terms assess competence on the same latent variable. To measure the same construct *across the whole lifespan* is a difficult endeavour as the NEPS, in contrast to many other educational studies, follows the development of persons from childhood to old age. Due to the long lifespan, the same tests may not be administered to the subjects, but tests need to be adapted in terms of difficulty and content to the relevant age group. This poses a challenge for the comparability of test results across time and across cohorts. Test scores need to be scaled on the same scale in order to ensure that differences in test scores can be attributed to differences in competencies and not to differences in the test. IRT provides means to make test scores of persons comparable, even when the persons did not receive the same test. However, certain test designs are necessary (e.g., common items or link studies, see explanation below) and the tests need to measure the same latent construct in different cohorts and ages. In the construction of test instruments a lot of effort is put on a coherent assessment of competencies over the lifespan (see, e.g., Gehrer et al., 2013, this issue; Neumann et al., 2013, this issue, or Hahn et al., 2013, this issue). For (almost) all age groups the same conceptual framework, the same cognitive demands, as well as the same item formats are used for test construction. Thus, from the construction point of view, the test developers aim at measuring the same la-

tent variable coherently across different age groups. If this effort proves successful, the data will enable researchers to investigate changes in mean and variance of the competence distributions across age as well as change, for example, exploring competence differences between two assessment waves or between different cohorts. Assuming that the construct measured is the same for different measurement occasions, a particular challenge of the NEPS is to equate results from consecutive assessments or different cohorts onto a common latent variable.

Within the framework of item response modeling, a number of models and methods have been discussed for linking and equating data from different assessments (for an overview see Kolen & Brennan, 2004; and von Davier & von Davier, 2007). One linking strategy refers to the measurement of mathematical competence in NEPS: Linking is performed by repeatedly presenting common items in consecutive assessments, thus facilitating the linking of these different tests via a "common-item nonequivalent groups design" (see, e.g., Kolen & Brennan, 2004; von Davier & von Davier, 2007). A similar linking model was used in the National Education Longitudinal Study of 1988 (NELS, Rock et al., 1995). An underlying assumption of this linking strategy is that the common items have the same item difficulty at both measurement occasions (measurement invariance). Measurement invariance will be investigated for all data collected. As retest effects are assumed for the measurement of reading competence and scientific literacy in the NEPS, a "common-item nonequivalent groups design" may not be applied. As an alternative linking strategy link samples are used instead. Items are presented to the subjects only once, and the link between two instruments for consecutive age groups relies on link samples (see von Davier & von Davier, 2007). In the NEPS, link samples are much smaller (N = 500 to N = 1,000) than main study samples (ranging between N = 3,000 to N = 15,000). They are randomly drawn from the elder of the two age groups, and the two test forms that are to be linked are administered to the link sample in a booklet design balancing for order effects within the assessment. Assumptions for this linking strategy are that measurement invariance across the main and the link sample holds and that both tests measure the same construct. The assumption of measurement invariance is very plausible for cohorts that are similar in age and situated in a similar institutional setting, such as linking Grade 5 to Grade 9 students, but it may be more questionable with respect to very different cohorts, such as Grade 9 students and adults. Adults differ from Grade 9 students not only in age (with a rather large age gap between both samples) but also in institutional setting (Grade 9 students being in school and used to tests and adults being part of the labor market). Thus, the assumption of measurement invariance is not a trivial one across such a long age span as is realized in NEPS. Whether this assumption holds true and whether both tests measure the same latent construct is investigated in the following study (see also Carstensen, Pohl, & Haberkorn, 2013).

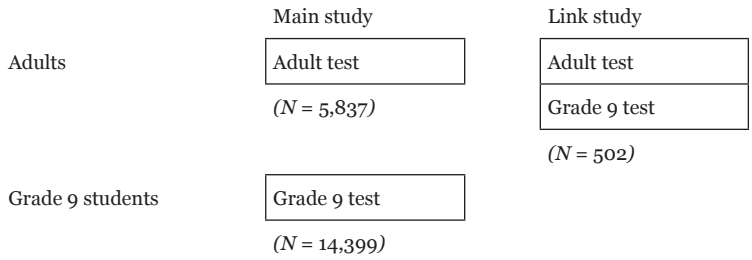
NEPS has already conducted some link studies and is planning to undertake more. By way of example we here present the results on measurement invariance for linking reading competence across Grade 9 students and adults. This study was

chosen because the large age gap and the differences in institutional setting pose specific challenges on measurement invariance, and while linking across school cohorts has already successfully be performed (e.g., in NELS, Rock, 2012), not many studies have worked on the problem of linking across such a long lifespan.

3.1 Method

The design for linking reading competence of Grade 9 students to reading competence of adults is shown in Figure 4. There are two main studies, that of Grade 9 students and that of adults. Both were conducted in 2010/2011 on a representative sample of Grade 9 students (N = 14,399) and adults (N = 5,837)[2], respectively, in Germany (Aßmann et al., 2011). The link study was conducted at the same time as the main study on N = 502 adults. A description of the three samples can be found in Table 2. The descriptive statistics on age, gender, migration background and books at home (as a proxy for socio-economic status) reveal that although, the sample of adults in the main study is quite similar to the sample of adults in the link study, there are some slight differences. The link sample is slighly younger, includes some more females and persons in that sample have less books at home than the persons in the main sample of adults. This may be due to the fact that persons in the main study agreed on participating in a panel study, while persons in the link study only took part once. Compared to students in Grade 9, there are fewer people in the link sample with a migration background and participants of the link study have slightly more books at home.

Figure 4: Design of the main study and the link study used for linking the Grade 9 reading test to the adult reading test



2 Note that due to data cleaning and data protection issues the number of participant may be slightly different from that in the Scientific Use File.

Table 2: Description of the three samples used for linking the Grade 9 reading test to the adult reading test

		Study		
Variable	Statistic	Main study adults	Link study adults	Main study Grade 9
Age	Mean	47.34	44.12	15.76
	SD	11.16	12.72	0.73
Gender	Rel. freq Female	50.9%	57.1%	49.4%
	Rel. freq Male	49.1%	42.5%	50.4%
Migration	Rel. freq No	79.9%	83.7%	69.4%
background	Rel. freq Yes	17.3%	15.5%	25.6%
	Rel. freq no information	2.9%	0.8%	5.1%
Books at home	Rel. freq 0–10	3.4%	6.6%	9.6%
	Rel. freq 11–25	6.8%	11.9%	13.0%
	Rel. freq 26–100	28.2%	31.4%	22.2%
	Rel. freq 101–200	20.7%	20.1%	20.2%
	Rel. freq 201–500	24.5%	16.9%	19.2%
	Rel. freq > 500	16.4%	12.9%	13.7%
	Rel. freq no information	0.0%	0.2%	2.1%

The reading tests that were linked across the two cohorts were developed by Gehrer et al. (2013, this issue). The items in the Grade 9 reading test differed from those for adults. There were no common items. However, the same conceptual framework has been used (i.e., five different text fuctions, different cognitive requirements, and different response formats; see Gehrer et al., 2013, this issue, for a description of the framework). While in the main studies, the Grade 9 students only received the Grade 9 reading test and the adults only received the adult reading test, both the Grade 9 and the adult reading test were administered to the link sample. In order to avoid order effects, the order of the test forms was randomly varied. Having access to data of the same subjects on both tests enables – if measurement invariance holds – the scaling of the two tests on the same scale. Whether the two tests really do measure the same construct was tested by investigating measurement invariance and dimensionality. Measurement invariance in this context means that the difficulty of the items is the same in the main sample and in the link sample and that subjects with the same reading competence have the same probability of correctly solving an item. Within the framework of IRT this is tested via DIF. DIF gives the difference in the estimated difficulty of the items for different samples (controlling for mean differences in ability). Dimensionality was checked by fitting a unidimensional and a two-dimensional model to the link sample. The two-dimensional model comprises the two different tests. To draw conclusions about dimensionality, we compared model fit and interpreted the cor-

relation between the two dimensions. If the two test forms measure the same latent construct, a unidimensional model should hold, the correlation between the two dimensions should be large (> .95, see Carstensen, 2013), and DIF should be small (<|.4| logits, see Pohl & Carstensen, 2012).

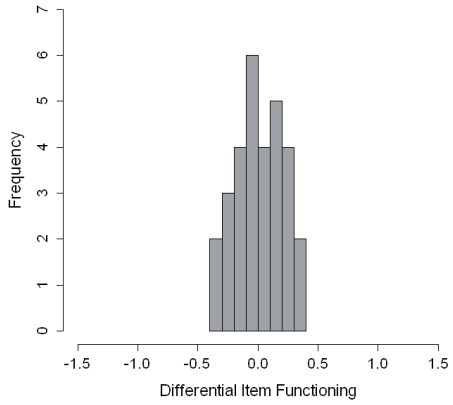
3.2 Results

Within the link sample we fitted a one-dimensional model (number of parameters = 82) across the items of both tests (Grade 9 and adult reading test) as well as a two-dimensional model (number of parameters = 84) separating the two tests. Both the AIC (1-dimensional model: 27667.0624, 2-dimensional model: 27655.9473) and the BIC (1-dimensional model: 28021.9876, 2-dimensional model: 28010.3097) favored the two-dimensional model. The correlation between the two dimensions (reading measured by the Grade 9 test and reading measured by the adult test) is, however, rather high (r = .939). There are some indications of multidimensionality of the two tests, that are, however, not severe. Regarding the high correlation, one could well argue for a unidimensional construct.

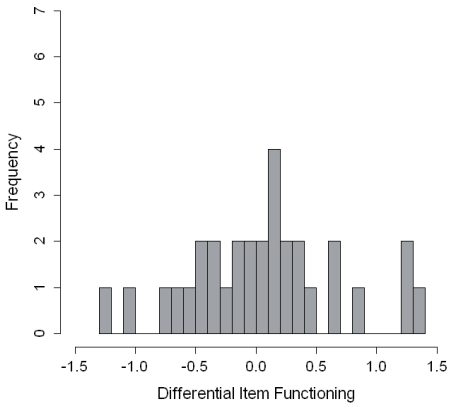
Measurement invariance was tested for the adult reading test, comparing the estimated item difficulties of the adult reading test in the adult main sample with those in the link sample. Measurement invariance for the Grade 9 test was investigated by comparing the estimated item difficulties of the Grade 9 reading test between the Grade 9 main sample and the link study (consisting of adults). The absolute differences in the estimated item difficulties for both tests are depicted in Figure 5. The results show that there is no considerable measurement invariance between the main sample and the link sample for the adult test. The largest difference in item difficulties is .34 logits. This is different for the Grade 9 test. Here, the largest difference in item difficulties between main sample and link sample is 1.4 logits. For the Grade 9 test, 14 out of 31 items show a considerable DIF of greater than .4 logits. The Grade 9 test measures a different construct when administered to adults than when administered to Grade 9 students. This result challenges the assumption of measurement invariance and, thus, also the construction of an appropriate link between these cohorts.

Figure 5: Differential item functioning comparing main and link study a) for the adult reading test and b) for the Grade 9 reading test

a) Adult test



b) Grade 9 test



3.3 Conclusion

In order to decide on an appropriate linking strategy, assumptions need to be tested. We thoroughly tested whether the Grade 9 reading test measured the same latent construct as the adult reading test. As both tests consist of distinct items, a linking study was conducted to allow for a concurrent scaling of the two tests. Dimensionality analyses supported the fact of a unidimensional reading construct. The results on DIF showed that measurement invariance held for the adult test, but not for the Grade 9 test. This might be due to the fact that the link sample consisted of adults. Grade 9 students and adults differ not only in age, but also in their institutional settings and most probably in their familiarity with tests. In fact, results on the number of missing values suggest that the different age groups used different test taking strategies. Whereas the pattern of missing values in the adult test is rather similar for the adults in the main study and the adults in the link study, it differs for the Grade 9 test between Grade 9 students in the main study and the adults in the link study. The adults in the link study have more omitted, not valid, and not reached items than the Grade 9 students in the main study. In further research reasons for impaired measurement invariance needs to be investigated.

The link study presented here was conducted on a sample of only one of two populations involved, namely, the adult population. Thus, we do not have any data from Grade 9 students working on the adult test instruments. This clearly is a restriction of our design however that is a result of practical constraints (resources). In order to derive a linking model, one can assume that the heterogeneity in item

difficulties and, as a quantification of this, the linking error obtained in the study conducted might be an estimate of the heterogeneity and linking error that would have been obtained in the linking study that was not conducted.

Thus, concurrent calibration (see, e.g., Kolen & Brennan, 2004; von Davier & von Davier, 2007), for which the same item difficulty is assumed in both samples, is not a suitable method for linking. Less restrictive linking methods establishing a link based on item parameters (e.g., a mean link) or on test information (e.g., Stocking-Lord method; see Kolen & Brennan, 2004) need to be considered. The rationale for less restrictive models may be that the particular items exhibit different difficulties due to the two populations compared, while at the same time the latent variable being assessed is the same. In further analyses these linking strategies will be applied and their suitability for the NEPS data will be evaluated.

In this paper, we have only regarded the link between Grade 9 students and adults. In future studies and analyses it will be interesting to see how well the assumptions of measurement invariance and unidimensionality are met for different age cohorts (e.g., kindergarten children and elementary school children) and for different competence domains (e.g., reading, mathematics, science, and also ICT) as well as across time (e.g., linking the same cohort across different measurement occasions).

4. Discussion

Large-scale studies, especially those with a longitudinal design pose specific challenges for the design and the scaling of competence data. Drawing on the data of the NEPS we have pointed out some of these challenges for which no satisfactory solution exist in the literature, yet and that have been approached in the NEPS. The challenges include topics regarding the scaling model (incorporating different response formats, dealing with missing values), design issues (adaptive testing in longitudinal designs), specific target groups, and aspects of linking. We have given an overview of research conducted in order to find appropriate answers to these questions and, more specifically, presented research conducted on the comparability of test scores across different age cohorts.

Although in longitudinal large-scale studies such as the NEPS numerous design aspects are challenging for finding appropriate testing and scaling procedures, the complex design and the large sample size also provide great opportunities for further development of methodological issues regarding competence testing and scaling. The results of the different studies presented in this paper, tackling different methodological challenges in large-scale studies, did not only serve to find appropriate testing and scaling procedures for the NEPS, but may also be relevant for other large-scale studies. For example, the incorporation of different response formats is also an issue in PISA. Different response formats are also present in other large-scale studies and the research approach as well as the results found in NEPS

may also be valid for studies such as PISA. The same holds true for the treatment of missing responses. So far, large-scale studies differ considerably in their approaches of dealing with missing values. Current research on the appropriateness of different methods has not yet been implemented in these large-scale assessments. Using the great data pool of the NEPS, which includes a wide range of age cohorts and different competence domains, we investigated whether assumptions of different approaches are met and have shown that ignoring missing responses is robust against violations of the nonignorability assumption. We are, thus, quite optimistic that the results found for NEPS may be generalized with respect to other low-stakes, large-scale assessments. In many large-scale studies (such as NAEP and PISA), great effort is put into the inclusion of students with SEN. No consistent conclusions can, as yet, be drawn as to whether and how this might be possible. With research being carried out in NEPS on the inclusion of students with SEN-L, we are adding to this discussion and are providing further arguments and solutions that may also be referred to by other large-scale studies. Finally, with longitudinal multi-stage testing, we specifically address practical constraints and the goal of an efficient and motivating testing design in longitudinal, large-scale studies. Longitudinal MST may also be combined with classical MST by routing not only within a test session (classical MST) but also to the first test form (longitudinal MST).

Although solutions could be found for many questions regarding the scaling of the NEPS competence data, a number of unresolved issues still remain. One of them is the estimation of plausible values that incorporate the complex design and are suitable for various kinds of research questions. As the released data will be used for a variety of research questions – which are unknown at the time of data release – providing appropriate plausible values for analyzing a broad range of research questions is a great challenge. Generally, the possible number of variables for the conditioning model is far too large for a reliable estimation of a conditioning model. The inclusion of various background variables in the measurement model for the estimation of plausible values will grow even more challenging with regard to the longitudinal design of the study as repeated measures from context variables will be available. Especially over a longer period of time, time-varying patterns in changes of context variables might be the focus of analyses and, thus, need to be incorporated in the conditioning model.

Another challenge arises from missing values in the conditioning variables. The missing responses have to be imputed and plausible values must be estimated, while still preserving the relationship of these variables and incorporating the uncertainty of imputation in the estimate of the plausible values. Other large-scale studies, such as PISA and NAEP, aggregate the questionnaire variables to orthogonal factors and use a set of factors (as many as needed to explain 90% of the variance of the questionnaire items) as background variables in the IRT measurement model of the competence data (Allen et al., 2001; OECD, 2009, 2012). Missing responses in background variables are single imputed, including other context variables, but not the competence score, in the imputation model. This approach is

a two-step approach that does not depict the uncertainty stemming from missing values in questionnaire items and does not account for the competence score in the imputation of missing responses on the context variables. Thus, point estimates of our research questions (e.g., regression coefficients, mean differences) as well as standard errors of these parameters may be biased. Current research in NEPS deals with this issue and a data-augmented Markov Chain Monte Carlo approach has been proposed (Aßmann, Carstensen, Gaasch, & Pohl, 2013) that would allow us to simultaneously impute the missing responses in the background variables and the estimation of plausible values. Following this approach, the latent competence score would be incorporated in the imputation model and the uncertainty of the imputation would be incorporated in the estimation of the plausible values. Before implementation in NEPS, however, this approach needs to be extended to incorporate larger sets of conditioning variables as well as multidimensional IRT models, and its performance needs to be compared to existing procedures.

Acknowledgments

This work was supported by third-party funds from the German Federal Ministry of Education and Research (BMBF). We thank Kerstin Haberkorn for her support in analyzing the data.

References

Abedi, J., Leon, S., Kao, J., Bayley, R., Ewers, N., Herman, J., & Mundhenk, K. (2011). *Accessible reading assessments for students with disabilities: The role of cognitive, grammatical, lexical, and textual/visual features* (CRESST Report 785). Los Angeles, CA: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.

Adams, R., & Wu, M. (Eds.). (2002). *PISA 2000* (*Technical Report)*. Paris, France: OECD Publishing.

Allen, N. A., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 Technical Report* (NCES 2001-452). Washington, DC: U.S. Department of Education.

Allen, N. L., Carson, J. E., Johnson, E. G., & Mislevy, R. J. (2001). Scaling procedures. In N. L. Allen, J. R. Donoghue, & T. L. Schoeps (Eds.), *The NAEP 1998 Technical Report* (pp. 227–246). Washington, DC: U.S. Department of Education.

Andrich, D. (1985). A latent trait model for items with response dependencies: Implications for test construction and analysis. In S. Embretson (Ed.), *Test design: Contributions from psychology, education and psychometrics* (pp. 245–273). New York, NY: Academic Press.

Angoff, W. H., & Huddleston, E. M. (1958). *A study of a two-level test system for the College Board Scholastic Aptitude Test* (Research Report SR-58-21). Princeton, NJ: Educational Testing Service.

Artelt, C., Weinert, S., & Carstensen, C. H. (2013). Assessing competencies across the lifespan within the German National Educational Panel Study (NEPS) – Editorial. *Journal for Educational Research Online 5*(2), 5–14.

Aßmann, C., Carstensen, C. H., Gaasch, J.-C., & Pohl, S. (2013). *Estimation of plausible values using partially missing background variables – A data augmented MCMC approach.* Manuscript submitted for publication.

Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., Rohwer, G., Rässler, S., & Blossfeld, H.-P. (2011). Sampling designs of the National Educational Panel Study: Challenges and solutions. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Education as a lifelong process. The German National Educational Panel Study (NEPS). Zeitschrift für Erziehungswissenschaft [Special Issue 14]* (pp. 51–65). Wiesbaden, Germany: VS.

Beard, J. J. (2008). *An investigation of vertical scaling with item response theory using a multistage testing framework* (Doctoral dissertation). Retrieved from http://ir.uiowa.edu/etd/216

Ben-Simon, A., Budescu, D. V., & Nevo, B. (1997). A comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement, 21*(1), 65–88.

Bielinski, J., Thurlow, M. L., Ysseldyke, J. E., Freidebach, J., & Freidebach, M. (2001). *Read-aloud accommodations: Effects on multiple-choice reading and math items* (NCEO Technical Report 31). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.). *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Blossfeld, H.-P., von Maurice, J., & Schneider, T. (2011). The National Educational Panel Study: Need, main feature, and research potential. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Education as a lifelong process. The German National Educational Panel Study (NEPS) [Special Issue 14]* (pp. 5–17). Wiesbaden, Germany: VS.

Bolt, S. E., & Ysseldyke, J. (2008). Accommodating students with disabilities in large-scale testing: A comparison of differential item functioning (DIF) identified across disability types. *Journal of Psychoeducational Assessment, 26*, 121–138.

Carstensen, C. H. (2013). Linking PISA competencies over three cycles – Results from Germany. In M. Prenzel, M. Kobarg, K. Schöps, & S. Rönnebeck (Eds.). *Research outcomes of the PISA Research Conference 2009* (p. 199–213). New York, NY: Springer.

Carstensen, C. H., Pohl, S., & Haberkorn, K. (2013, March). *Längsschnittliche Kalibrierung im Nationalen Bildungspanel am Beispiel von Lesekompetenz und Mathematikkompetenz.* Presentation at the 1st meeting of the Gesellschaft für Empirische Bildungsforschung, Kiel, Germany.

Cormier, D. C., Altman, J., Shyyan, V., & Thurlow, M. L. (2010). *A summary of the research on the effects of test accommodations: 2007-2008* (Technical Report 56). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana, IL: University of Illinois Press.

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.

Duchhardt, C., & Gerdes, A. (2012). *NEPS Technical Report for Mathematics – Scaling results of Starting Cohort 3 in fifth grade* (NEPS Working Paper No. 19). Bamberg, Germany: Otto-Friedrich-Universität, Nationales Bildungspanel.

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

Finch, H., Barton, K., & Meyer, P. (2009). Differential item functioning analysis for accommodated versus nonaccommodated students. *Educational Assessment, 14*, 38–56.

Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2013). NEPS framework for assessing reading competence *and* results from an adult pilot study. *Journal for Educational Research Online, 5*(2), 50–79.

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

Haberkorn, K., Pohl, S., Carstensen, C. H., & Wiegand, E. (2013). *Incorporating different response formats of NEPS competence tests in an IRT model*. Manuscript submitted for publication.

Hahn, I., Schöps, K., Rönnebeck, S., Martensen, M., Hansen, S., Saß, S., & Dalehefte, I. M., & Prenzel, M. (2013). Assessing science literacy over the lifespan – A description of the NEPS science framework and the test development. *Journal for Educational Research Online, 5*(2), 110–138.

Heydrich, J., Weinert, S., Nusser, L., Artelt, C., & Carstensen, C. (2013). Including students with special educational needs into large-scale assessments of competencies: Challenges and approaches within the German National Educational Panel Study (NEPS). *Journal for Educational Research Online, 5*(2), 217–240.

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.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NJ: Springer.

Koretz, D. M. (1997). *The assessment of students with disabilities in Kentucky* (CSE Technical Report 431). Los Angeles, CA: CRESST/RAND Institute on Education and Training.

Koretz, D. M., & Barton, K. (2003). *Assessing students with disabilities: Issues and evidence* (CSE Technical Report 587). Los Angeles, CA: University of California, Center for the Study of Evaluation. Retrieved from http://www.cse.ucla.edu/products/reports/TR587.pdf

Linn, R. L., Rock., D. A., & Cleary, T. A. (1969). The development and evaluation of several programmed testing methods. *Educational and Psychological Measurement, 29*, 129–146.

Lord, F. M. (1971a). A theoretical study of two-stage testing. *Psychometrika, 36*, 227–242.

Lord, F. M. (1971b). The self-scoring flexilevel test. *Journal of Educational Measurement, 8*, 147–151.

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

Lovett, B. J. (2010). Extended time testing accommodations for students with disabilities: Answers to five fundamental questions. *Review of Educational Research, 80*, 611–638.

Lukhele, R., & Sireci, S. G. (1995). *Using IRT to combine multiple-choice and free-response sections of a test on to a common scale using a priori weights*. Paper presented at the annual conference of the National Council on Measurement in Education, San Francisco, CA.

Lutkus, A. D., Mazzeo, J., Zhang, J., & Jerry, L. (2004). *Including special-needs students in the NAEP 1998 reading assessment* (ETS Research Report ETS-NAEP 04-R01). Princeton, NJ: Educational Testing Service.

Macaskill, G., Adams, R. J., & Wu, M. L. (1998). Scaling methodology and procedures for the mathematics and science competence, advanced mathematics and physics scale. In M. Martin & D. L. Kelly (Eds.). *Third International Mathematics*

and Science Study, Technical Report Volume 3: Implementation and analysis. Chestnut Hill, MA: Boston College.

Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (Eds.). (2007). *PIRLS 2006 Technical Report*. Chestnut Hill, MA: Boston College.

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

McKinley, R. L., & Reckase, M. D. (1980). Computer applications to ability testing. *Association for Educational Data Systems Journal, 13*, 193–203.

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.

Mislevy, R. J., & Wu, P.-K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (RR-98-30-ONR). Princeton, NJ: Educational Testing Service.

Moustaki, I., & Knott, M. (2000). Weighting for item non-response in attitude scales by using latent variable models with covariates. *Journal of the Royal Statistical Society, Series A: Statistics in Society, 163*, 445–459.

Mullis, I. V. S., Martin, M. O., & Diaconu, D. (2004). Item analysis and review. In M. O. Martin, I. V. S. Mullis, & S. J. Chrostowski (Eds.), *TIMSS 2003 Technical Report*. (pp. 225–252). Chestnut Hill, MA: Boston College.

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

Neumann, I., Duchhardt, C., Ehmke, T., Grüßing, M., Heinze, A., & Knopp, E. (2013). Modeling and assessing mathematical competence over the lifespan. *Journal for Educational Research Online, 5*(2), 80–109.

OECD – Organisation for Economic Co-operation and Development. (2009). *PISA 2006 Technical Report*. Paris, France: OECD Publishing.

OECD – Organisation for Economic Co-operation and Development. (2012). *PISA 2009 Technical Report*. Paris, France: OECD Publishing.

Olson, J. F., Martin, M. O., & Mullis, I. V. S. (Eds.). (2008). *TIMSS 2007 Technical Report*. Chestnut Hill, MA: Boston College.

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: Statistics in Society, 162*, 177–194.

Pearson. (2011, October 11). Pearson to develop framework for OECD's PISA students assessment for 2015. Retrieved from http://www.pearson.com/media-1/announcements/?i=1485

Pitoniak, M. J., & Royer, J. M. (2001). Testing accommodations for examinees with disabilities: A review of psychometric, legal, and social policy issues. *Review of Educational Research, 71*, 53–104.

Pohl, S. (in press). Longitudinal multi-stage testing. *Journal of Educational Measurement*.

Pohl, S., & Carstensen, C. (2012). *NEPS Technical Report – Scaling the data of the competence tests* (NEPS Working Paper No. 14). Bamberg, Germany: Otto-Friedrich-Universität, Nationales Bildungspanel.

Pohl, S., Gräfe, L., & Rose, N. (in press). Dealing with omitted and not reached items in competence tests – Evaluating approaches accounting for missing responses in IRT models. *Educational and Psychological Measurement*.

Pohl, S., Haberkorn, K., Hardt, K., & Wiegand, E. (2012). *NEPS Technical Report for reading – Scaling results of starting cohort 3 in fifth grade* (NEPS Working Paper No. 15). Bamberg, Germany: Otto-Friedrich-Universität, Nationales Bildungspanel.

Pohl, S., Hardt, K., Südkamp, A., Nusser, L., & Weinert, S. (2013). *Testing special educational needs students in large-scale-assessments: Identifying assessable students with Mixed-Rasch-Hybrid models*. Manuscript in preparation.

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rost, J., & Schiefele, U. (Eds.). (2006). *PISA 2003. Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres*. Münster, Germany: Waxmann.

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

Rock, D. A. (2012). *Modeling change in large-scale longitudinal studies of educational growth: Four decades of contributions to the assessment of educational growth*. (ETS Research Report No. RR-12-04), Princeton, NJ: Educational Testing Service.

Rock, D. A., Pollack, J. M., & Quinn, P. (1995). *National Education Longitudinal Study of 1988. Psychometric Report for the NELS:88 Base Year Through Second Follow-Up* (Rep. No. NCES 95-382). Washington, DC: U.S. Department of Education.

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT)* (ETS Research Report No. RR-10-11). Princeton, NJ: Educational Testing Service.

Senkbeil, M., Ihme, J. M., & Wittwer, J. (2013). The test of Technological and Information Literacy (TILT) in the National Educational Panel Study: Development, empirical testing, and evidence for validity. *Journal for Educational Research Online, 5*(2), 139–161.

Si, C. B. (2002). *Ability estimation under different item parameterization and scoring models*. Denton, TX: University of North Texas.

Sireci, S. G., Scarpati, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75*, 457–490.

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

Südkamp, A., Pohl, S., & Weinert, S. (2013). *Assessing reading comprehension of students with special educational needs – Identification of appropriate testing accommodations*. Manuscript submitted for publication.

Sykes, R. C., & Hou, L. (2003). Weighting constructed-response items in IRT-based exams. *Applied Measurement in Education, 16*, 257–275.

von Davier, M., & von Davier, A. A. (2007). A unified approach to IRT scale linking and scale transformations. *Methodology, 3*, 115–124.

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Boston, MA: Kluwer.

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. New York, NY: Springer.

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. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

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.

Warm, T. A. (1989). Weighted likelihood estimation of ability in Item Response Theory. *Psychometrika, 4,* 427–450.

Wilson, M. (1988). Detecting and interpreting local item dependence using a family of Rasch models. *Applied Psychological Measurement, 12*, 353–364.

Wilson M., & Adams R. J. (1995). Rasch models for item bundles. *Psychometrika, 60*, 181–198.

Wongwiwatthananukit, S., Bennett, D. E., & Popovich N. G. (2000). Assessing pharmacy student knowledge on multiple-choice examinations using partial-credit scoring of combined-response multiple-choice items. *American Journal of Pharmaceutical Education, 64*, 1–10.

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

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest Version 2.0: Generalised item response modelling software*. Camberwell, Australia: Australian Council for Educational Research.

Xu, X., Sikali, E., Oranje, A., & Kulick, E. (2011, April). *Multi-stage testing in educational survey assessments*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME) in New Orleans, LA.

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, CO.