Modeling Competence Data in Large-Scale Educational Assessments

Inaugural-Dissertation

in der Fakultät Humanwissenschaften

der Otto-Friedrich-Universität Bamberg

vorgelegt von

Kerstin Haberkorn

aus Tirschenreuth

Bamberg, den 03.02.16

Tag der mündlichen Prüfung: 17.12.2015

Dekan:	Prof. Dr. Stefan Hörmann, Otto-Friedrich-Universität Bamberg
Erstgutachter:	Prof. Dr. Claus H. Carstensen, Otto-Friedrich-Universität Bamberg
Zweitgutachter:	Prof. Dr. Steffi Pohl, Freie Universität Berlin

Danksagung

Ich möchte mich an dieser Stelle bei verschiedenen Personen bedanken, die mich bei der Erstellung dieser Arbeit begleitet und unterstützt haben.

Zuerst möchte ich mich bei Steffi Pohl für die intensive Betreuung meiner Dissertation und für Ihre tolle Mentorentätigkeit bedanken. Durch sie konnte ich zahlreiche wissenschaftliche und nicht-wissenschaftliche Fähigkeiten weiterentwickeln. Sie hat die wertvolle Eigenschaft, in wissenschaftlichen Arbeiten und der beruflichen Tätigkeit immer wieder den „roten Faden“ zu suchen und zu besprechen. Ich danke ihr für den wertschätzenden und aufmerksamen Austausch!

Daneben bedanke ich mich ganz herzlich bei Claus H. Carstensen, meinem Promotionsvater, der immer ein offenes Ohr für meine Fragen hatte und stets konstruktiv zum Gelingen der Arbeit beigetragen hat. Gerade als ich die Promotion nebenberuflich weitergeführt habe, konnte ich mich auf seine gute Erreichbarkeit und unkomplizierte, schnelle Kommunikationswege verlassen, was maßgeblich zur Realisierbarkeit der Promotion beigetragen hat.

Bei Kathrin Lockl möchte ich mich ebenfalls bedanken. Sie war gerade für meinen Artikel zum metakognitiven Wissen eine wichtige und kompetente Ansprechpartnerin. Von ihren inhaltlichen Ideen und ihrem Blick für interessante Fragestellungen habe ich sehr profitiert.

Ein besonderer Dank gilt meinen Freundinnen und Freunden für willkommene Ablenkungen und beständiges Nachfragen zum Stand meiner Dissertation! Insbesondere möchte ich Anna und Sebastian für regelmäßige Bibliotheksbesuche und leckere Mittags- und Kaffeepausen danken, Katja und Carmen für anregende Diskussionen, Angelika für selbstgesetzte Deadlines zum eigenen Ansporn und Verena für ihren herrlichen Optimismus und den besten Ressourcen-Blick, den man sich vorstellen kann.

Meiner Familie, besonders meinen Eltern Angela und Franz, möchte ich ein großes Dankeschön aussprechen für den bereichernden telefonischen und persönlichen Austausch in allen Phasen der Promotion. Ihr großes Vertrauen in meine Fähigkeiten hat mir auch in schwierigen Phasen der Promotion wertvollen Rückenwind gegeben!

Und schließlich möchte ich mich bei meinem Freund Markus für sein riesiges Verständnis und seine Unterstützung bedanken! Für ihn war es selbstverständlich, dass ich während der letzten Jahre so viel Zeit und Energie in die Promotion verwendet habe. Ich danke ihm für seine Geduld und sein tolles Rahmenprogramm während meiner Promotion!

Contents

Abstract

Over the last decades, large-scale assessments focusing on skills and knowledge of individuals have expanded significantly in order to obtain information about conditions for and consequences of competence acquisition. To provide valid and accurate scores of the subjects under investigation, it is necessary to thoroughly check the psychometric properties of the competence instruments that are administered and to choose appropriate scaling models for the competence data. In this thesis, various challenges in modeling competence data were addressed that arose from different recently developed competence tests in large-scale assessments. The different tests posed specific demands on the scaling of the data such as dealing with multidimensionality, incorporating different response formats, or linking competence scores. By investigating these challenges associated with each of the competence tests, the aim of the thesis was to draw implications for the specification of the scaling models for the competence data.

First, a new metacognitive knowledge test for early elementary school children was investigated. As earlier findings on metacognitive knowledge in secondary school pointed to empirically distinguishable components of metacognitive knowledge, especially the dimensionality of the newly developed test was studied. Therefore, uni- and multidimensional models were applied to the competence data and their model fit was compared. By applying multidimensional latent-change models, the homogeneity of change was observed as further indicator for dimensionality. Overall, the new test instrument exhibited good psychometric properties including fairness of the items for various subgroups. In accordance with previous studies in other age groups the results indicated a multidimensional structure of the newly developed test instrument. In the discussion, theoretical as well as empirical arguments were compiled that should be considered for the choice of a uni- or a multidimensional model for the metacognitive knowledge data.

The next objective in the thesis was to study a series of reading competence tests intended to measure the same latent trait across a large age span. The different reading competence tests, developed in a longitudinal large-scale study, were based on the same conceptual framework and were administered from fifth grade to adulthood. We specifically investigated whether the test scores were comparable across such a large age span enabling to interpret change across time. The analyses on the reading competence tests showed that the coherence of measurement could not fully be assured across the wide age range. The application of strict linking models allowing for the interpretation of developmental progress seemed to be justified within secondary school, but not between secondary school and adulthood.

The last purpose in the thesis was to find out how to adequately incorporate different response formats in a scaling model. Therefore, multiple choice (MC) and complex multiple choice (CMC) items were regarded as they are most frequently used in large-scale assessments. Specifically, we explored whether the two response formats form distinct empirical dimensions and which a priori scoring schemes for the two response formats appropriately model the competence data. The results demonstrated that the response formats built a unidimensional measure across domains, studies, and age cohorts justifying to use a unidimensional scale score. A differentiated scoring of the CMC items yielded a better discrimination between persons and was, thus, preferred. The a priori weighting scheme of giving each subtask of a CMC item half the weight of a MC item described the empirical competence data well.

1 Synopsis

1.1 Introduction

As the interest in educational processes and their impact on individual life courses as well as on economic growth has increased within information society, large-scale assessments collecting competence data across nations and time have expanded considerably over the last years (Blossfeld, Schneider, & Doll, 2009; Kirsch, Lennon, von Davier, Gonzalez, & Yamamoto, 2013). The systematic surveys on students' educational attainment provide information for a variety of stakeholders, such as policymakers, economists, school principals, teachers and social scientists in the respective countries (Hanushek & Woessmann, 2008, 2011; Ritzen, 2013). The large-scale studies usually assess a variety of competence domains as well as background variables about educational institutions and the private life. Thus, a broad range of questions concerning competence development, influencing factors and consequences can be addressed. Furthermore, relevant conclusions on further educational policies may be derived from the results obtained by the educational assessments. In order to draw adequate inferences from the educational data on the underlying trait, it is crucial to thoroughly check the competence tests' quality and to develop adequate scaling models for analyzing the competence data.

The thesis aimed to shed light on some important questions arising in the context of an appropriate modeling of large-scale competence data. How may multidimensionality adequately be dealt with in scaling the competence data? How may different measurement occasions be implemented? How may items with different response formats be treated in the scaling model? In the following sections, at first a short overview is given on general psychometric models applied to competence data in large-scale assessments. Then, relevant aspects of test quality are presented and implications for the specification of scaling or linking models are delineated. References to

each of the four manuscripts are given which focused on specific challenges associated with competence test data. Note that the descriptions of psychometric models and empirical test properties are not intended to be exhaustive, but to provide a brief overview and to point out specific issues that were of particular relevance in the present thesis. Afterwards, the research questions and results of each manuscript are detailed.

1.2 Modeling Competence Data in Educational Assessments

In educational assessments which focus on the acquirement of competencies, often new competence tests are designed and, then, administered to the desired sample. Before drawing inferences from the competence data about the individuals' knowledge, a convenient psychometric model for scaling the data should be chosen and the functioning of the test should be evaluated (Rost, 2004). According to Wilsons' framework of construct modeling (2005) the choice and evaluation of a measurement model and the evaluation of a test's reliability and validity can be summarized as quality control methods after installing the test instrument. Each of the quality control steps may provide information about the test instrument, the underlying construct and the appropriateness of the scaling model. Following, valuable information may be obtained for specifying the final scaling model. In the thesis, the different newly developed competence tests all posed specific demands such as multidimensionality or measurement variance. Therefore, at first their test quality was thoroughly examined and, secondly, implications for the final scaling model were delineated.

The choice of the *general* psychometric model and the tests' evaluation are closely related. Usually a psychometric model for scaling the data is selected, the fit of the model to the specific data is examined and, in consequence, the test instrument and/or the final scaling model is adjusted. Whereas in the last decades several guidelines have been published focusing on a thorough construction of test instruments (see, e.g., Downing & Haladyna, 2006; Haladyna & Rodriguez, 2013; Osterlind, 1998), less attention has been paid to detailed quality checks of competence tests in educational settings and to a deliberate representation of test characteristics in the scaling model.

1.2.1 The Choice of a General Scaling Model

The model that relates the outcomes of a test back to the construct is often termed the measurement model or the psychometric model (Wilson, 2005). For educational large-scale data, Item response theory (IRT) models have become state of the art because of their flexibility and their great potential in solving measurement problems (Embretson & Reise, 2000). In response to specific needs of the competence assessment, several IRT models have been developed and existing have been modified and extended recently. One of the most relevant models developed in the context of IRT is the one parameter (1PL) model (Rasch, 1960[1]). The 1PL model models the distance between person locations and item locations in a probability function enabling to place persons and items on the same scale. Each item is characterized by one parameter, the item location parameter. Many large-scale assessments make use of a 1PL type model with a constant slope parameter and location parameters for each item such as the English Language Proficiency Assessment (ELPA; e.g., Council of Chief State School Officers, 2012), the Program for International Student Assessment (PISA; e.g., OECD, 2013) or the National Educational Panel Study (NEPS; e.g., Blossfeld, von Maurice, & Schneider, 2011). Other popular IRT models are extensions of the 1PL model such as the two parameter (2PL) logistic model or the three parameter (3PL) logistic model. Whereas the one parameter model (1PL) assumes that all items have the same item discrimination, in the two parameter (2PL) model an additional discrimination parameter is introduced to model the deviances in discrimination based on the item's empirical capacity to differentiate among the subjects' abilities (de Ayala, 2009; Embretson & Reise, 2000). The 3PL model additionally includes a guessing parameter. The 2PL model is, for instance, applied in the Adult Literacy and Lifeskills survey (ALL; e.g., OECD &

[1] There are slight differences between the Rasch model and the 1PL model (see, for instance, de Ayala, 2009). The slope constant in the Rasch model is 1.0, whereas in the 1PL model the slope constant has not to be equal to 1.0. For simplicity reasons, in the following the term 1PL model is used including all models with a constant value for the item discrimination parameter.

Statistics Canada, 2005) or in the National Assessment of Educational Progress (NAEP; Jones & Olkin, 2004), the 3PL model is applied in the Trends in International Mathematics and Science Study (TIMSS; e.g., Mullis, Martin, Foy, & Arora, 2012) or in the Progress in International Reading Literacy Study (PIRLS; e.g., Mullis, Martin, Foy, & Drucker, 2012).

As argued in Manuscript 4 of the thesis, the choice of the general psychometric model for scaling the competence data primarily depends on theoretical deliberations. The different proposed IRT models all have their advantages and their limitations. The 2PL (and the 3PL) model have the possibility to represent competence data in more detail by containing two (or three) varying parameters. Thus, a better fit of the measurement model to the competence data is obtained in comparison to a 1PL model. The 1PL model, in contrast, is more sparse, and, thus, may underestimate the variances of discrimination in the items. However, violations of the equal slopes assumption seem not to strongly bias the ability and difficulty estimates in a 1PL model (Forsyth, Saisangjan, & Gilmer, 1981; Wainer & Wright, 1980). An advantage of the 1PL model is that it allows for implementing theoretical considerations about the weighting of items. Since the weight of the items on the overall competence score is modeled only by the a priori scoring of the items, the item weights can be determined deliberately based on the theoretical framework of the test developers.

Having selected the general scaling model for the educational data, the empirical properties of the competence test are investigated. As shown in the manuscripts of the thesis, it is crucial to consider and evaluate different theoretical assumptions made in the measurement model, such as dimensionality of subcomponents of the construct or theoretically delineated a priori weighting schemes in the process of checking test quality. The results may provide valuable information on the test instrument and – if the test was thoroughly constructed – on the underlying construct or

the response formats, respectively. As a consequence, questions concerning the further specification of the scaling model can be addressed: Should the final scaling model be uni- or multidimensional? How may different measurement occasions be implemented? How may different response formats be scored appropriately?

1.2.2 Psychometric Properties of the Test Instrument and the Specification of the Scaling Model

Tests instruments may be characterized by a variety of empirical criteria such as item difficulty, test targeting, item fit, dimensionality, measurement invariance, reliability, and so on (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; Pohl & Carstensen, 2012; Wilson, 2005). The section especially focuses on item fit, the internal structure of the competence test, and measurement invariance across groups and time, as these properties provided specific challenges in the competence tests analyzed in the present thesis. For each of the psychometric properties, relevant analyses to investigate them and the relation to the choice and specification of the scaling model are shown. Specific issues investigated by the four manuscripts of the thesis are illustrated. Note that in the following section the approaches applied in the manuscripts and general implications for scaling models are described. The specific results obtained from the analyses of the thesis are detailed in Section 1.3.

1.2.2.1 Item Fit

Usually, various measures are drawn on for evaluating the item fit (OECD, 2014; Pohl & Carstensen, 2012). On the one hand, the empirical item characteristic curves are studied. The aim is to detect items which have a flat or non-increasing curve, as these items show an inconsistent performance with the underlying model. Furthermore, correlations between the item score and

the total score and between the distractors and the total score are examined. In the 2- and 3PL model, the discrimination parameter is regarded that characterizes how well the item discriminates among the examinees. A low discrimination parameter suggests that the item does not differentiate well between respondents and, thus, that the item does not provide great information about the examinees (Embretson & Reise, 2000). In the 1PL model other fit criteria are examined, as the 1PL model does not estimate different discrimination parameters. In order to evaluate if the items fulfill the assumption of an equal discrimination among the items, several fit indices have been developed (Andersen, 1973; Glas, 1988; Wright & Masters, 1982). Commonly, the observed and the model expected residuals for responses on the items are compared producing chi-square-like statistics to detect low fitting items. An often used criterion to describe the degree of deviation between the probabilities is, for instance, the (weighted) mean square fit statistic by Wright and Masters (1982). Analyzing the competence tests of the thesis, always item fit analyses were performed as basic analyses for checking the quality of the items.

Implications of the Results for the Specification of the Scaling Model

The results of the item fit statistics depend on the underlying psychometric model and may also affect its specification. Overall, a number of items showing a dissatisfying fit to the 1PL model indicate that the assumption of a common discrimination parameter across items is challenged. In this case, excluding items and/or developing new items or including additional parameters in the model might be ways to obtain a better fit of the measurement model to the competence data. Additionally, as shown in the third and fourth manuscript of the thesis, item fit statistics may be useful to evaluate a priori scoring schemes for different item types in order to deliberately specify the final scaling model. In the third manuscript of the thesis, item fit measures were employed among other measures to evaluate different scoring rules for CMC items. Specifically, the

appropriateness of common procedures of aggregating response categories of polytomous items was evaluated. Based on the results, recommendations for the aggregation of polytomous items in scaling models for competence data were derived. In the fourth manuscript of the thesis, different a priori weighting schemes were compared in order to find out how adequately they represent the empirical competence data. The results provided evidence that some of the a priori weighting schemes yielded a considerable misfit of the different response formats, whereas one a priori weighting scheme reflected the competence data well across studies, domains, and age groups. Several implications for an appropriate implementation of different response formats in a scaling model were drawn in the manuscript. First, during the process of checking test quality response formats might be more likely to be retained if their a priori weights do not well describe the empirical information obtained by them. Thus, the impact of a priori weights on the item fit should be considered when evaluating different item types. Second, it might always be useful to evaluate a priori considered weights for different response formats for choosing an a priori scoring scheme that appropriately reflects the amount of information the response formats carry.

Altogether, basic psychometric characteristics of the items may not only be used to optimize items or select items for the final scaling. As shown in the manuscripts, they may also be valuable to evaluate and reconsider a priori scoring or weighting schemes and enable researchers to establish a deliberate final scaling model.

1.2.2.2 Dimensionality

The 1-, 2-, or 3PL models assume a unidimensional continuum of the ability measured by the competence tests. However, competence tests often do not completely meet this assumption. Therefore, it should be empirically tested whether the assumed unidimensionality holds or whether different components assessed in a competence test prove to be empirically

distinguishable. Overall, there are different sources which may lead to multidimensionality of a competence test[2] (for an overview, see Embretson & Reise, 2000; Reckase, 2009). Frequently, competence tests are constructed to catch a broad and comprehensive construct including several subdimensions. In educational settings, competence instruments are often designed based on a conceptual framework. In PISA or NEPS, for instance, scientific literacy tests comprise items testing knowledge of science and items targeted towards knowledge about science (Hahn et al., 2013; OECD, 2013). The subcomponents of a construct – although intended to be unidimensional - might empirically form a multidimensional structure and, thus, challenge building a unidimensional competence score. In Manuscript 1 of the thesis, unidimensionality of the newly developed test on metacognitive knowledge was also challenged. Several analyses were performed to investigate whether the test instrument comprises distinct metacognitive knowledge components. Specifically, it was examined whether the metacognitive knowledge data was better represented by a model based on different strategy dimensions, a model based on different mental processes or a model assuming a unidimensional latent trait. Overall, the results provided valuable insight into the structure of the metacognitive knowledge test and also delivered relevant information for specifying the final scaling model. Another source that may result in multidimensionality are different response formats included in a competence test. Different response formats may demand different cognitive processes from the examinees yielding multidimensionality (Ackerman & Smith, 1988; Palmer & Devitt, 2007). Unfortunately, dimensionality of response formats in competence tests is only rarely tested so far. In Manuscript 4, dimensionality of the two most common response formats in large-scale studies was studied. The overall purpose was to find out, how to adequately incorporate MC and CMC items in a

[2] In the following, the description focuses on between-item multidimensionality (see, for instance, Adams, Wilson, & Wang, 1997). A short outlook on other multidimensional IRT models is given in the discussion.

scaling model. Therefore, it was also explored whether MC and CMC items empirically form a multidimensional or a unidimensional measure.

The assumption of unidimensionality can be tested by applying uni- and multidimensional models to the competence data. Model fit criteria of the uni- and the multidimensional models are used that indicate how well the models describe the data. First, likelihood ratio tests for nested models (the unidimensional model can be seen as special case of the multidimensional model) are drawn on that show if the multidimensional model fits the data significantly better than the unidimensional model. Additionally, information criterion indices of the models such as Akaikes (1974) Information Criterion (AIC) or the Bayes Information Criterion (BIC; Schwarz, 1978) varying in their strength to penalize additional parameters may be compared. Moreover, the correlations between the latent dimensions constituted by the subfacets of a construct or the different response formats are usually examined that exhibit how closely the different components are related. The different criteria were all used in the thesis for evaluating the research questions concerning dimensionality. In the first manuscript of the thesis, not only correlations between the subcomponents at the first and the second measurement point were regarded, but a further indicator for dimensionality of a construct was investigated. An only recently developed method by Vautier and Pohl (2009) was adapted to explore the homogeneity of change. Therefore, the dimensionality of the latent change between the different strategy dimensions was examined between Measurement Point 1 and 2. The aim of these analyses was to find out whether changes between the subdimensions were highly related indicating a homogeneous evolvement and providing further evidence for unidimensionality.

Implications of the Results for the Specification of the Scaling Model

When the test instrument encompasses a broad and representative sample of items measuring the underlying construct, the analyses on dimensionality may shed light on the structure of the latent trait or provide evidence about the functioning of the response formats, respectively. Furthermore, the results of checking a tests' dimensionality may affect the specification of the scaling model. As described in Manuscript 1, evidence for multidimensionality raises the question whether a uni- or a multidimensional model should be used for scaling the competence data (as usually a unidimensional latent space was intended). In the manuscript it is recommended to take both theoretical and statistical reasons into account for specifying the scaling model. From a statistical perspective, applying unidimensional models to a multidimensional test assessing different subfacets of a construct might bias the results. According to Walker and Beretvas (2003), a higher standard error of the ability estimates and the proficiency classification might result from violations to the unidimensionality assumption. However, Reckase (2009) emphasized to also take the parsimony criterion into account. He stated that the use of more complex models is only justified when they yield an increased accuracy or when new insights are gained by them. Thus, when differences between the unidimensional and the multidimensional model in terms of their fit to the data are small, the unidimensional model might be preferred because of parsimony reasons. Some model fit measures, for instance, the BIC already comprise the parsimony criterion by penalizing for additional parameters in the statistical model to account for both fit and complexity of the model. From a theoretical perspective, it might enrich research to analyze components of a multidimensional test separately regarding their development and the relation to other variables to get a more detailed and accurate picture of the construct. Yet, it may also be argued that the focus of interest is to find out more about a heterogeneous construct in its entirety. The operationalization might thus include heterogeneous

aspects of the trait to broadly represent the construct of interest. Nevertheless, an overall estimate of the respective competence might then be justified to approximate the latent trait (Ercikan, 2006).

1.2.2.3 Measurement Invariance

In educational research, investigations often address the comparison of different populations or intend to assess development of a competence. An important condition for such a comparison between groups or across time is measurement invariance between the different subpopulations and measurement points that are investigated. Measurement invariance has been stated as crucial test criterion by the American Educational Research Association (American Educational Research Association et al., 2014), as fundamental comparisons in large-scale studies, for instance, between countries, schools, or grades, may only be valid when the measurements are equivalent across groups. Thus, the prominence of checking measurement invariance of competence tests increased constantly in the last years. Measurement invariance implies that differences in an assessment between persons or groups are due to their latent trait and not due to a different functioning of items for a particular group of respondents (Millsap, 2010; Widaman & Reise, 1997). In the first manuscript, measurement invariance of the metacognitive knowledge test was checked for relevant subgroups based on gender, design (longitudinal vs. cross-sectional), and migration background. Furthermore, measurement invariance across time from first to second grade was investigated to justify that the same latent variable was measured between the two measurement occasions. In the second manuscript, the issue of measurement invariance was addressed across a wide age span. In order to test for the coherence of measurement in reading competence, a series of reading competence tests administered from fifth grade to adulthood were analyzed.

To test measurement invariance of a test instrument, researchers usually work within the IRT framework, since IRT enables to produce sample invariant statistics. Specifically, researchers explore whether the relationship of the items with the latent variable is the same across subgroups. Violations of measurement invariance are termed item bias or differential item functioning (DIF). One popular approach of examining DIF is to apply multiple group models to the data, separately estimate item parameters for the subgroups and compare them (for a review on examining DIF, see Millsap & Everson, 1993). Overall group differences or differences in standard deviations are not signs of DIF, but they have to be accounted for when examining DIF. In the meantime, several rules of thumbs have been proposed for classifying the size of DIF (Zwick, Thayer, & Lewis, 1999; Pohl & Carstensen, 2012; OECD, 2013). When analyzing measurement invariance longitudinally, the term Item Parameter Drift (IPD) is used to denote items which show differences in item difficulties in different waves of assessment after controlling for overall group differences (Holland & Wainer, 1993). In Manuscript 1, analyses on item parameter drift were performed between Grade 1 and 2. Therefore, a model with constraint item parameters between the two measurement occasions was applied and the estimated item parameters were compared with a model without constraints on item parameters. In Manuscript 2, specific link studies implemented in the NEPS enabled to check measurement invariance across test forms and age groups. In addition to competence assessments in Grade 5, 9, and adults, additional samples in the link studies had taken two competence tests of adjacent age groups. To find out whether the reading competence construct remained the same across test forms, dimensionality of the different reading competence tests was investigated within the link samples. To find out whether the reading competence construct changed across samples, measurement invariance of the same reading test was examined across age groups.

Implications of the Results for the Specification of the Scaling Model

Results concerning the measurement invariance might be of interest for re-inspecting the items' content, searching for reasons for DIF or item parameter drift, and for refining or removing the items. If measurement invariance does not hold between subsamples of the study, substantive group comparisons are challenged. However, if there are comprehensible arguments that explain threats to measurement invariance, but do not challenge validity of the assessed construct within a subsample (e.g., items of a depression scale which function differently in a group of depressed students and non-depressed students) within a subsample, it might be promising to analyze the subgroups and relationships to other variables separately. In vertical assessments of competencies, lacks in measurement invariance may affect the choice of the linking model used for comparing the different points in time. In Manuscript 2, different ways of dealing with violations of measurement invariance in the linking model are discussed. Overall, there are a number of models for establishing a vertical scale (Camilli, Yamamoto, & Wang, 1993; Williams, Pommerich, & Thissen, 1998; Yen, 1986). When there is only a small amount of items showing item parameter drift, a strict linking strategy may be applied to the competence tests data with restrictions on item difficulty on an item level. This linking strategy may also allow for interpreting the competencies' trajectories over time. When there are a number of items exhibiting item parameter drift, a less restricted linking with a larger link error might be more appropriate. These models may be useful for gaining first impressions about differences in age cohorts, while taking into account that the constructs that were assessed are not exactly comparable. In Manuscript 1, the analyses on IPD exhibited that measurement invariance between Grade 1 and 2 did not hold for few items of the metacognitive knowledge test. Therefore, the analyses of the latent change of the subdimensions were performed applying a

model with partial measurement invariance. All items were implemented in the model, but equality constraints between grades were only posed for 11 of the 14 items.

1.3 Manuscripts of This Thesis

As mentioned in the previous section, the thesis comprises four manuscripts. In this section, title and references of the manuscripts are presented, the research questions are detailed and a summary of the results is given.

As each of the manuscript addressed research questions that were of particular importance for the respective domain or the response format, the manuscripts do not show all steps of checking empirical properties of the test instrument. Assuming that the test instrument included a representative set of items for the respective domain, implications from the results for (a) the test instrument (b) the underlying construct, and (c) the scaling model can be derived. In the first section of the thesis, we described relevant empirical properties of competence tests and implications for specifying the scaling model with references to the manuscripts of the thesis. In this section, the research questions that were examined in the four manuscripts are specified and the results of the analyses are briefly presented with respect to the different levels of implications (a-c).

1.3.1 Manuscript I: Dimensionality of a New Metacognitive Knowledge Test

Haberkorn, K., Lockl, K., Pohl, S., Weinert, S., & Ebert, S. (2014). Metacognitive knowledge in children at early elementary school. *Metacognition and Learning*, *9,* 239-263.

Summary

Knowledge about mental processes and strategies is a central factor for successful learning in institutional contexts as well as in out-of-school environments. However, there was a lack on group tests assessing metacognitive knowledge in early elementary school economically and validly. Therefore, a new test on children's metacognitive knowledge had been developed in the

BiKS-3-10 study on Educational Processes, Competence Development, and Selection Decisions at Preschool and Elementary School Age. Previous research provided evidence that test instruments assessing metacognitive knowledge might form a multidimensional structure. However, researchers had not yet examined whether components of metacognitive knowledge were empirically distinguishable in young school children. Therefore, the purpose of the first manuscript of the thesis was to thoroughly evaluate the new test instrument on metacognitive knowledge and, in particular, examine the dimensionality of the metacognitive knowledge test. The test instrument was administered to children at the end of first grade and one year later at the end of second grade. For the analyses of the dimensionality of the metacognitive knowledge test, not only multidimensional models within one measurement occasion were applied to the data. Also the heterogeneity of change in the underlying components of the construct was studied as indicator for dimensionality. Overall, 14 out of 15 items exhibited good psychometric properties and the reliability of the test instrument was acceptable. The results of the differential item functioning (DIF) analyses of the items indicated that measurement invariance was ensured across the relevant variables gender, design (longitudinal vs. cross-sectional design) as well as migration background. In first as well as in second grade evidence occurred that the subdimensions in the metacognitive knowledge test did not fully measure the same latent trait. Nevertheless, the change of the dimensions form first to second grade in the children was rather homogeneous supporting the assumption for unidimensionality. The discussion of the manuscript addressed the issue of dealing with multidimensional competence data when an overall score for the competence test was intended.

1.3.2 Manuscript II: Linking Reading Competence Tests Across the Life Span

Pohl, S., Haberkorn, K., & Carstensen, C. (in press). Measuring competencies across the lifespan – challenges of linking test scores. In M. Stemmler, A. von Eye, & W. Wiedermann (Eds.). *Dependent data in social sciences research: Forms, issues, and methods of analysis.* Springer.

Summary

In the context of large-scale studies, there is growing interest in the assessment of competencies across time in order to examine change of the competencies within the subjects and compare different age cohorts. However, several assumptions need to hold for making meaningful comparisons across measurement occasions and cohorts. Overall, the measurement of the competence that is assessed in the different age groups needs to be coherent. Specifically, the tests administered to the different age groups must be measurement invariant. Additionally, different reading competence tests administered to the participants across age must measure the same construct. So far, measurement invariance has only been investigated across small age ranges. The objective of the second manuscript was, thus, to investigate whether a coherent measurement of competencies may also be obtained across a large age span. For our analyses, we drew on data of the NEPS, as the NEPS – in contrast to many other large-scale studies – considers competence development across the whole life span. Specifically, we focused on a series of reading competence tests from the NEPS administered from Grade 5 to adulthood. As retest effects were assumed for reading competence in the NEPS, additional link studies had been performed to link the different starting cohorts of the NEPS in Grade 5, Grade 9, and adults. In the link studies the two competence tests of adjacent age groups had been administered. In order to check for the comparability of the competencies, we explored unidimensionality of the test forms as well as measurement invariance across age groups. The results provided evidence that

the measures of reading competence were unidimensional within the link samples. However, the differential item functioning analyses showed that measurement invariance was only present across the school cohorts, but not between Grade 9 students and the adult sample. The differences of these two cohorts in age, in the institutional setting, and in competence levels seemed to yield differences in the functioning of the items. In the discussion of the manuscript, possible reasons for measurement invariance such as different missing processes were illustrated and implications of measurement variance for linking the different cohorts were delineated.

1.3.3 Manuscript III: Aggregation of Complex Multiple Choice Items

Haberkorn, K., Pohl, S., Carstensen, C., & Wiegand, E. (in press). *Scoring of complex multiple choice items in NEPS competence tests.* In H.-P. Blossfeld, J. von Maurice, M. Bayer, & J. Skopek (Eds.). *Methodological issues in longitudinal surveys.* Springer.

Summary

Usually, competence tests consist of a variety of response formats to adequately and validly measure the participants' knowledge. The most widely used item types in educational assessments are MC and CMC items. Multiple choice items contain an item stem and different response options with one of them being correct. Complex multiple choice items comprise a number of dichotomous true-false items. Whereas there is consensus among researchers that MC items are given one point when answered correctly and zero points otherwise, the scoring of CMC items varies across studies. Usually, the subtasks of CMC items are aggregated to a polytomous variable, but different approaches of aggregating the categories of the polytomous items exist. So far, these aggregation options have rarely been investigated using IRT. The third manuscript of the thesis focused on examining which of different aggregation options for CMC items appropriately models the empirical competence data. One of the common aggregation

options for CMC items is the All-or-Nothing scoring rule. The All-or-Nothing scoring implies that participants only receive full credit if all subtasks are solved, otherwise they get no credit. Another common scoring option is the Number Correct scoring rule. This scoring rule means that subjects receive partial credit for each correctly solved subtask. Using ICT and science competence tests from the NEPS, we compared the effects of the two aggregation options on item difficulty, discrimination, reliability parameters and the range of person abilities within categories that were collapsed. The results showed consistently that a considerable amount of information is lost by applying the All-or-Nothing scoring rule. Therefore, the use of a differentiated scoring without an aggregation of categories of the polytomous CMC items was recommended in the discussion to best discriminate between subjects under investigation.

1.3.4 Manuscript IV: Dimensionality and Weighting of Multiple Choice and Complex Multiple Choice Items

Haberkorn, K., Pohl, S., & Carstensen, C. (2015). Incorporating different response formats of competence tests in an IRT model. Manuscript submitted for publication.

Summary

Associated with competence tests embedding MC and CMC items, further questions concerning the implementation of the two response formats in a scaling model arise: Do the MC and CMC items measure the same latent trait and, if so, which impact should the two response formats have on the overall competence score? Should they be weighted equally? Or should CMC items comprising more subtasks contribute more to the overall competence score? So far, results on dimensionality concerning MC and CMC item types have been limited and not fully consistent. Moreover, different a priori weighting schemes for MC and CMC items have been applied to competence tests using 1PL models. Yet, it has not been investigated how appropriately these

weighting schemes describe the empirical competence data. We, thus, thoroughly addressed the research questions concerning dimensionality and weighting of the two response formats by examining a variety of competence data. In order to delineate meaningful implications on how the two response formats can be treated adequately in a scaling model, we analyzed different domains (ICT, Science), different studies (NEPS, PISA), and different grades (G6, G9). Overall, the two item types empirically formed a unidimensional structure in all competence tests justifying the construction of a unidimensional scale score. Furthermore, the a priori weighting scheme of giving half the weight of MC items to subtasks of CMC items appropriately reflected the empirical weight of the MC response formats in comparison to the CMC response formats. Implications of the results for the implementation of MC and CMC items in a scaling model were drawn.

1.4 Discussion

1.4.1 Integrating the Research Findings

Overall, the purpose of the thesis was to address specific demands that were posed by different newly developed competence tests. Thorough analyses provided valuable information about the test instruments and conclusions for appropriately including the specific test characteristics into the scaling of the competence data could be derived.

Regarding the challenges involved in scaling the test instruments of the thesis, an overall concern in modeling competence data may be seen in an appropriate dealing with heterogeneity. Heterogeneity in competence data can arise from a heterogeneous set of items included in the test instrument as well as from a heterogeneous sample of subjects participating in the study. Reckase (2009) described a conceptual framework in which he considered differences in the subjects taking a test and the items constituting a test instrument as factors that – together - create a multidimensional space. According to Reckase, the interaction of the persons and items is affected by the number of dimensions of variability in the participants that complete a test and the number of dimensions of sensitivity of the test items. With regard to the manuscripts of the thesis, the research on the test instrument assessing metacognitive knowledge in young children and on different response formats in particular focused on investigating the heterogeneity of the items and ways to adequately implement their characteristics in a scaling model. In the case of the metacognitive knowledge test, different classification systems for the items existed. The analyses on the metacognitive knowledge test aimed at empirically comparing these classification systems, examining their appropriateness for describing the empirical data and delineating implications for the scaling model. In the NEPS competence tests, different response formats were examined as possible sources yielding multidimensionality of the items. Furthermore, the

weights of the different response formats were explored as another facet of heterogeneity of the items in addition to multidimensionality. Ways of appropriately implementing the different item types in a scaling model were discussed. In the research on linking competencies across the lifespan, finally, both heterogeneity of items and persons were investigated as factors that might lead to a multidimensional continuum. Both assumptions, that is, unidimensionality of the items as well as unidimensionality of the samples were checked to answer the question whether meaningful comparisons across age cohorts may be drawn. As only the prerequisite of a unidimensional space across the test forms was fulfilled, but not across the samples, ways to deal with the heterogeneity across age groups for linking competence data were illustrated in the discussion.

1.4.2 Implications for the Specification of Scaling Models

In the following, at first strengths and limitations of the current research are discussed with regard to the specification of scaling models in the context of large-scale studies. An outlook on further research that might be promising in this area is provided. Afterwards, the results of the thesis are discussed with regard to the implications for the specific contents and the response formats, respectively.

1.4.2.1 Strengths and Limitations of the Research

One strength of the thesis is that different issues in the field of an appropriate modeling of competence data were considered which had received only little attention so far in educational large-scale studies. Many large-scale assessments aim at following developmental trajectories of competencies. However, it has not yet been explored whether assumptions for linking test scores across a large age span hold empirically. Furthermore, it is still not state of the art to evaluate whether different response formats included in a test instrument form a unidimensional construct.

Whereas dimensionality of competence tests comprising different subdimensions is, by now, usually explored by educational researchers in the process of checking test quality (see, e.g., Haberkorn, Pohl, Hardt, & Wiegand, 2012; OECD, 2014), different response formats are usually analyzed together and composite scores are built across all items. At last, a few large-scale studies make use of the 1PL model implying an a priori weighting of different item types. These a priori weighting schemes have only rarely been evaluated. Nevertheless, a priori weights which differ considerably from empirical weights of the response formats may bias conclusions about item types. The manuscripts in the thesis attempted to analyze these important questions in the establishment of a scaling model and delineate recommendations for researchers working with competence data. The thesis was, thus, aimed at making a valuable contribution to guidelines for a deliberate modeling of competence data.

Other strengths of the thesis are the broad data base and the elaborate analyses that were conducted to examine the research questions. The questions concerning dimensionality, weighting, or measurement invariance of the test instruments were always addressed by investigating a number of studies or measurement occasions. Furthermore, different IRT analyses were conducted and a variety of statistical criteria were investigated for each of the research issues. In the process of evaluating the metacognitive knowledge test developed for early elementary school children, dimensionality in first as well as in second grade was studied in order to draw conclusions about the heterogeneity of the items of the test instrument. Furthermore, a sophisticated multidimensional latent-change model adopted from Vautier and Pohl (2009) was applied to the data to not only evaluate the dimensionality separately in first as well in second grade, but also examine change in the subdimensions as indicator for dimensionality. To investigate whether the assumptions for linking competencies across the lifespan were reached, a series of link studies were analyzed and not only data of reading competence, but also data of

mathematical competence was drawn on. The research on a priori weighting schemes for different response formats finally comprised analyses on different grades (G6, G9), domains (science, ICT), and studies (NEPS; PISA) for checking the generalizability of the results. Moreover, the a priori weighting schemes were not only evaluated by applying simple PC models but also by applying newly developed restricted 2PPC models. In sum, the thorough and wide range of analyses build a strong basis for the implications derived from the results. Additionally, the analyses in the thesis may well be used in other large-scale studies for investigating dimensionality, assumptions for linking test scores, or a priori weighting schemes in order to specify the final scaling model.

The presented research has some limitations concerning the specification of scaling models. In the manuscript on metacognitive knowledge of elementary school children and the manuscript on linking test scores, the focus was set to specific challenges of the test instruments such as dimensionality of the items or measurement invariance across samples that have to be investigated in order to appropriately model the respective competence data. However, implications for the scaling models were only described, but the scaling models were not developed and tested in detail. Furthermore, the thesis provided only snapshots about relevant issues that should be taken into account in the specification of a scaling model. Of course, there are many other challenges one has to meet in scaling competence data. These include questions on how to deal with missing values, how to implement tests administered in different positions in the booklet, or how to estimate unbiased population estimates. Finally, the present thesis made especially use of the 1PL model or extensions of the 1PL model, for instance the partial credit model. Actually, some of the research questions only arise when using 1PL models such as the comparison of a priori weighting schemes for different response formats. Hence, researchers preferring the application of 2- or 3PL models will probably deal differently with some of the

research issues presented in the thesis. However, challenges such as multidimensionality across items or samples seem to be of great relevance for scaling competence data independent of the specific measurement model (1PL, 2PL, or 3PL model…) that is applied.

1.4.2.2 Outlook on Future Research

One future research task should be to evaluate in more detail the performance of psychometric models that were considered in the discussion sections of Manuscript 1 and 2. Concerning the dimensionality of competence tests, different multidimensional models have been introduced in the last decades to model complex domains and multiple abilities (Adams, Wilson, & Wang, 1997; McDonald, 2000; Reckase, 2009). The different models for multidimensional data differ in their complexity and, thus, in their closeness to represent the empirical data. Some of these models allow for modeling a multidimensional structure and, yet, forming an overall competence score across items (Wang & Wilson, 2005). With regard to the metacognitive knowledge test, the different models may be applied and empirical differences in the parameter estimates may be compared. Furthermore, the multidimensional models may be compared to less complex models to broaden findings on the robustness of more parsimonious models to the multidimensional structure. A few simulation studies, for instance, suggest that unidimensional IRT models may be robust to moderate degrees of multidimensionality (Ackerman, 1989; Kirisci, Hsu, & Yu, 2001). Concerning the linking of competencies across time, a variety of IRT methods such as concurrent calibration or fixed parameters scale linking exist (see, e.g., Kolen & Brennan, 2004; Von Davier, Carstensen, & von Davier, 2008) and restrictions on item difficulty on item level or on test level are possible. Depending of the degree of violation to the comparability assumptions, appropriate models should be applied to the NEPS reading competence tests. The empirical performance of the approaches should be investigated with respect to item and person parameter estimates and linking errors should be compared.

Since a broad range of questions has to be answered for the specification of scaling models for competence data, further research is needed to approach other challenges surrounded with educational competence tests. In large-scale studies usually time-limited competence tests are administered which may yield a non-negligible amount of missing responses. These missing values have to be adequately accounted for, and, hence, research is necessary to investigate how the missing responses can be treated correctly in the scaling model (Holman & Glas, 2005; Köhler, Pohl, & Carstensen, 2014; Pohl, Gräfe, & Rose, 2014). Another challenge for scaling competence data are different kinds of multidimensional structures that have to be evaluated and implemented appropriately in the scaling model. Whereas the thesis focused on multiple dimensions in a test with items referring to one of the dimensions (between-item multidimensionality), multidimensional structures may also appear by multiple abilities assessed by one item (within-item multidimensionality). Scaling models may be developed that account for these relationships and that are practical in empirical assessments (Embretson, 1984; Walker & Beretvas, 2001, 2003). Further challenges in scaling large-scale competence data arise from implementing background variables in the measurement model and estimating plausible values. Future research might find adequate ways of incorporating relevant background variables including time-varying background information for providing population estimates for a variety of research questions.

1.4.3 Implications for the Respective Domains and Response Formats

1.4.3.1 Strengths and Limitations of the Research

In the following, for each domain and response format, respectively, the main strengths and limitations of the results obtained in the thesis are briefly summarized.

The investigation of metacognitive knowledge in elementary school children provided valuable insight about the mental processes and strategies the children acquire throughout their first elementary school years. A main strength of the manuscript for the research on metacognitive knowledge seems to be that for the first time dimensionality of a metacognitive knowledge test for elementary school children was thoroughly analyzed. The results corroborated theoretical assumptions about the heterogeneity of metacognitive knowledge (Flavell & Wellman, 1977) and findings from secondary school children (Neuenhaus, Artelt, Lingel, & Schneider, 2011). Of course, dimensionality of a test depends on the development and the selection of items for the test instrument. Although a broad and representative sample of items was included in the test, the number of items within the dimensions was restricted due to motivation and time limit reasons. Thus, conclusions on the underlying construct are still limited.

A particular strength of the study investigating linking reading competence with regard to the content was that not only assumptions for linking across the samples were tested, but also reasons for measurement variance between students in school and adults at work were studied. No relationship between DIF of items and text functions or cognitive requirements could be found. However, valuable insights into test-taking behavior could be obtained by comparing the missingness patterns between students in Grade 9 and the sample of adults. Shortcomings of the study were that the examination of differences between the samples was limited to the information available from the large-scale data set. Relevant variables such as gender or migration background, and additionally the occurrence of missing responses could be compared between the samples, but, no cognitive interviews (see, for instance, Prüfer & Rexroth, 2005) for information about cognitive operations during the tasks were available due to the large-scale setting. Furthermore, as the tests were administered in paper-and-pencil mode, no information about the exact response times of the participants were available to explore the possible

differences in the response processes between the G9 students and adults in more detail. Thus, the conclusions about factors that might lead to measurement variance between the samples were limited.

A strength of the analyses of different response formats was that the unidimensionality of MC and CMC items was checked for the ICT as well as for the science domain and was found in the PISA as well as in the NEPS survey. Hence, the findings corroborated and enlarged previous research using a second language ability test (Dudley, 2006) and a medical achievement test (Downing, Baranowski, Grosso, & Norcini, 1995). Furthermore, also the cognitive processes accompanied with the response formats were reviewed and the empirical results of the study were compared with theoretical considerations about the response formats. Limitations of the study with regard to implications for the response formats were again that no precise conclusions on the mental operations activated by the response formats could be derived. The results suggested that similar processes may be involved in answering the different item types yielding no source for multidimensionality, but the specific cognitive processes were not identified in the study.

1.4.3.2 Outlook on Future Research

In the area of metacognitive knowledge of young children, further empirical research should focus on the evolvement of the different components of metacognitive knowledge and differential relations to other competence domains. As the thesis shed light on the heterogeneity of the construct, the impact of single dimensions of metacognitive knowledge on school performance might be of interest for researchers as well as for classroom teachers (Artelt, Schiefele, & Schneider, 2001; Veenman, Kok, & Blöte, 2005). Moreover, it might be valuable to implement the investigation of dimensionality as standard procedure when evaluating competence tests. So far, these analyses are primarily conducted in the context of large-scale assessments, but are still

not state of the art in smaller educational studies, e.g. in projects investigating metacognition of children.

Concerning the linking of reading competence, future research should explore in more detail the test-taking behavior across samples. Since first evidence was provided that the Grade 9 students and adults differed considerably in their missingness patterns, a thorough investigation of response times and skipping of items might be promising (Zerpa, Hachey, van Barnfield, & Simon, 2011). Furthermore, the content of items exhibiting a large item drift should be analyzed. Though no relations of item drift and cognitive requirements or text functions were found, there might be other relevant features of the items which may lead to item drift. Finally, research on other domains across this age-span might show whether violations of the prerequisites for linking occur rather domain-specific or domain-independent across samples. First analyses indicate that threats to measurement invariance are also present for mathematical competence.

Future research on the functioning of MC and CMC items should be concerned with the empirical validation of the specific cognitive processes associated with the two response formats. So far, empirical studies have revealed unidimensionality of the MC and CMC response format across domains. However, empirical results concerning the intellectual processes that are involved in solving the tasks are still rare. For reading comprehension, van den Bergh (1990) found that processes of recall and recognition are present in answering MC items. Further analyses on MC and CMC items in other domains are needed to derive cognitive models about the answering process of these item formats. Additionally, further studies should investigate the functioning of the two response formats in samples of adults. So far, research on MC and CMC items has concentrated on participants at school or at university who are familiar with different response formats in competence tests (Dudley, 2006; Frisbie & Sweeney, 1982).

2 References

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13,* 113-127.

Ackerman, T. A., & Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. *Applied Psychological Measurement, 12,* 117-128.

Adams, R.,Wilson, M., &Wang,W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1-23.

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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Education Research Association.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38,* 123-140.

Artelt, C., Schiefele, U. & Schneider, W. (2001). Predictors of reading literacy. *European Journal of Psychology of Education, 16*, 363–383.

Blossfeld, H.-P., Schneider, T., & Doll, J. (2009). Die Längsschnittstudie Nationales Bildungspanel: Notwendigkeit, Grundzüge und Analysepotential. *Pädagogische Rundschau, 63*, 249-259.

Blossfeld, H.-P., von Maurice, J., & Schneider, T. (2011). The National Educational Panel Study: Need, main features, and research potential. *Zeitschrift für Erziehungswissenschaft, 14*, 5-17.

Camilli, G., Yamamoto, K., & Wang, M. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement, 17*, 379-388.

Council of Chief State School Officers (CCSSO). (2012). *Framework for english language proficiency development standards corresponding to the Common Core State Standards and the Next Generation Science Standards.* Washington, DC: Author.

De Ayala, R. J. (2009). *The theory and practice of item response theory.* New York, NY: Guilford Press.

Downing, S. M., & Haladyna, T. M. (Eds.). (2006). *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum.

Downing, S. M., Baranowski, R. A., Grosso, L. J., & Norcini, J. J. (1995). Item type and cognitive ability measured: The validity evidence for multiple true-false items in medical specialty certification. *Applied Measurement in Education, 8,* 187-197.

Dudley, A. (2006). Multiple dichotomous-scored items in second language testing: investigating the multiple true-false item type under norm-referenced conditions. *Language Testing, 23,* 198-228.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika, 49,* 175-186.

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists.* Mahwah: Erlbaum Publishers.

Ercikan, K. (2006). Developments in assessment of student learning and achievement. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (pp. 929–953). Mahwah, NJ: Lawrence Erlbaum.

Flavell, J. H., & Wellman, H. M. (1977). Metamemory. In R. V. Kail & J. W. Hagen (Eds.), *Perspectives on the development of memory and cognition* (pp. 3–34). Hillsdale, NJ: Lawrence Erlbaum.

Forsyth, R., Saisangjan, U., & Gilmer, J. (1981). Some empirical results related to the robustness of the Rasch model. *Applied Psychological Measurement, 5,* 175-186.

Frisbie, D. A., & Sweeney, D. C. (1982). The relative merits of multiple true–false tests. *Journal of Educational Measurement, 19*, 29-35.

Glas, C. A. W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika, 53,* 525-546.

Haberkorn, K., Pohl, S., Hardt, K., & Wiegand, E. (2012). *NEPS technical report for reading – Scaling results of starting cohort 4 in ninth grade (NEPS Working Paper No. 16)*. Bamberg, Germany: University of Bamberg, National Educational Panel Study.

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

Haladyna, T. M., & Rodriguez, M. C. (2013) *Developing and validating test items*. New York, NY: Routledge.

Hanushek, E. A., & Woessmann, L. (2008). The role of cognitive skills in economic development. *Journal of Economic Literature, 46,* 607-668.

Hanushek, E. A., & Woessmann, L. (2011). How much do educational outcomes matter in OECD countries? *Economic policy, 26,* 427-491.

Holland, P. W. & Wainer, H. (Eds.). (1993). *Differential item functioning: Theory and practice.* Hillsdale, N.J.: Erlbaum.

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.

Jones, L. V. & Olkin, I. (Eds.). (2004). *The nation's report card: Evolution and perspectives.* Bloomington, IN: Phi Delta Kappa Educational Foundation.

Kirisci, L., Hsu, T.-C., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement, 25,* 146-162.

Kirsch, I., Lennon, M., von Davier, M., Gonzalez, E., & Yamamoto, K. (2013). On the growing importance of international large-scale assessments. In M. von Davier, E. Gonzales, I. Kirsch & K. Yamamoto (Eds.). *The role of international large-scale assessments: Perspectives from technology, economy and educational research* (pp. 1-11). Dordrecht, Netherlands: Springer.

Köhler, C., Pohl, S., & Carstensen, C. H. (2014). Taking the missing propensity into account when estimating competence scores: Evaluation of item response theory models for nonignorable omissions. *Educational and Psychological Measurement, 1,* doi: 10.1177/0013164414561785.

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

McDonald, R.P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement, 24,* 99-114.

Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data: An introduction. *Child Development Perspectives, 4,* 5-9.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 297-334.

Mullis, I. V., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V., Martin, M. O., Foy, P., & Drucker, K. T. (2012). *PIRLS 2011 international results in reading.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Neuenhaus, N., Artelt, C., Lingel, K., & Schneider,W. (2011). Fifth graders metacognitive knowledge: general or domain specific? *European Journal of Psychology of Education, 26*, 163–178. doi:10.1007/s10212–010–0040–7.

OECD & Statistics Canada (2005). *Learning a living: First results of the adult literacy and life skills survey.* Paris, France: OECD.

OECD (2013*). PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy.* Paris, France: OECD.

OECD (2014). *PISA 2012 technical report*. Paris, France: OECD.

Osterlind, S. J. (1998*). Constructing test items: Multiple-choice, constructed-response, performance, and other formats.* Dordrecht, Netherlands: Kluwer Academic.

Palmer, E. J., & Devitt, P. G. (2007). Assessment of higher order cognitive skills in undergraduate education. Modified essay or multiple-choice questions. *BMC Medical Education, 7,* 49. Retrieved from http://www.biomedcentral.com/1472-6920/7/49/

Pohl, S., & Carstensen, C. H. (2012). *NEPS technical report – Scaling the data of the competence tests.* (NEPS Working Paper No. 14). Bamberg, Germany: University of Bamberg, National Educational Panel Study.

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement, 74,* 423-452.

Prüfer, P., & Rexroth, M. (2005). Kognitive Interviews. *ZUMA How-to-Reihe*, Nr. 15, 1-21.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Nielsen & Lydiche.

Reckase, M. D. (2009). *Multidimensional item response theory.* New York, NY: Springer.

Ritzen, J. (2013). International large-scale assessments as change agents. In M. von Davier, E. Gonzales, I. Kirsch & K. Yamamoto (Eds.). *The role of international large-scale assessments: Perspectives from technology, economy and educational research* (pp. 13-24). Dordrecht, Netherlands: Springer.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* (2nd ed.). Bern, Switzerland: Hans Huber.

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

van den Bergh, H. (1990). On the construct validity of multiple-choice items for reading comprehension. *Applied Psychological Measurement, 14,* 1-12.

Vautier, S., & Pohl, S. (2009). Bipolarity of latent change in STAI scores. *Psychological Assessment, 21,* 187-193.

Veenman , M. , Kok , R. , & Blöte , A. (2005). The relation between intellectual and metacognitive skills in early adolescence. *Instructional Science, 33,* 193 – 211.

von Davier, A. A., Carstensen, C. H., & von Davier, M. (2008). Linking competencies in horizontal, vertical and longitudinal settings and measuring growth. In J. Hartig, E. Klieme & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 121-149). New York, NY: Hogrefe & Huber.

Wainer, H., & Wright, B. D. (1980). Robust estimation of ability in the Rasch model. *Psychometrika, 45,* 373-391.

Walker, C. M., & Beretvas, S. N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: A cognitive explanation for DIF. *Journal of Educational Measurement, 38,* 147-163.

Walker, C. M., & Beretvas, S. N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement, 40*(3), 255–275.

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

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariances of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle & S. G. West (Eds.), *The science of prevention: methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.

Williams, V. S. L., Pommerich, M., & Thissen, D. (1998). A comparison of developmental scales based on Thurstone methods and item response theory. *Journal of Educational Measurement, 35*, 93-107.

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

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

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23*, 299-325.

Zerpa, C., Hachey, K., van Barnfield, C., & Simon, M. (2011). Modeling student motivation and students' ability estimates from a large-scale assessment of mathematics. *SAGE Open, 1,* 1–9.

Zwick, R., Thayer, D., & Lewis, C. (1999). An empirical bayes approach to mantel-haenszel DIF analysis. *Journal of Educational Measurement, 36*, 1-28.

3 Appendix

3.1 Manuscript 1: Dimensionality of a New Metacognitive Knowledge Test

Original Source of Publication

Haberkorn, K., Lockl, K., Pohl, S., Ebert, S., & Weinert, S. (2014). Metacognitive knowledge in children at early elementary school. *Metacognition and Learning, 9*(3), 239-263, doi:10.1007/s11409-014-9115-1

Metacognition Learning (2014) 9:239–263
DOI 10.1007/s11409-014-9115-1

Metacognitive knowledge in children at early elementary school

Kerstin Haberkorn · Kathrin Lockl · Steffi Pohl · Susanne Ebert · Sabine Weinert

Received: 30 April 2013 / Accepted: 23 April 2014 /
Published online: 21 May 2014

Abstract In metacognition research, many studies focused on metacognitive knowledge of preschoolers or children at the end of elementary school or secondary school, but investigations of children starting elementary school are quite limited. The present study, thus, took a closer look at children's knowledge about mental processes and strategies in early elementary school aiming to extend findings on the respective age period. Therefore, at first, a new test that can be administered in group settings and that assesses a broad concept of children's metacognitive knowledge in early elementary school was evaluated. Furthermore, analyses on the structure of metacognitive knowledge were carried out in cross-sectional as well as longitudinal analyses. In a longitudinal design, the new test instrument was administered to 870 children at the end of first grade and again one year later (N=720). Item Response models were used to evaluate the construct validity of the test. Test characteristics were checked based on different fit statistics, test fairness, and discriminant validity. In summary, the test exhibited good psychometric properties. Analyses on the dimensionality of the assessed metacognitive knowledge revealed that different strategies seemed to form rather distinct dimensions of metacognitive knowledge. However, these dimensions showed a rather homogeneous development from first to second grade. Impacts of the findings on theoretical considerations and on the theoretical understanding of metacognitive knowledge and further analyses with metacognitive competence data are discussed.

Keywords Metacognitive knowledge · Test evaluation · Structure · Longitudinal design

K. Haberkorn (✉) · K. Lockl
Leibniz Institute for Educational Trajectories, University of Bamberg, Wilhelmsplatz 3, 96047 Bamberg, Germany
e-mail: kerstin.haberkorn@lifbi.de

S. Pohl
Department of Methods and Evaluation/Quality Management, Free University of Berlin, Habelschwerdter Allee 45, 14195 Berlin, Germany

S. Ebert · S. Weinert
Department of Developmental and Educational Psychology, University of Bamberg, Markusplatz 3, 96047 Bamberg, Germany

Introduction

Over the last few decades, there has been considerable interest in children's knowledge about their memory, learning, and comprehension processes (e.g., Artelt et al. 2001; Brown et al. 1983; Flavell et al. 2002; Joyner and Kurtz-Costes 1997). Within the broader construct of metacognition, metacognitive knowledge has been conceptualized as the declarative component besides children's procedural activities in regulating and monitoring memory performance during a task (Flavell 1979; Schneider and Pressley 1997; Veenman et al. 2006). Metacognitive knowledge can be broken down into knowledge about the different mental processes of remembering, understanding, and learning (Annevirta and Vauras 2001). It can also be subdivided according to strategies that are involved in learning situations, for example, organizational or text processing strategies (Schlagmüller et al. 2001). The relevance of metacognitive knowledge for memory functioning has been shown in many studies (e.g., DeMarie et al. 2004; Justice 1985; Schneider and Bjorklund 1998; Weinert and Schneider 1999). The researchers provided evidence that metacognitive knowledge contributes substantially to children's use of strategies in task performance. Schlagmüller et al. (2001) finally confirmed that the connection between metacognitive knowledge and memory functioning is mediated by children's effective application of strategies during memory tasks. Apart from that, children's knowledge about their mental activities has been studied in its own right as a central aspect of their cognition about the world (Flavell et al. 2002).

In the field of metacognitive research, a number of studies focused on metacognitive development in early childhood and researchers especially examined preschoolers' knowledge about memory activities and its precursors (Lockl and Schneider 2006, 2007; O'Sullivan 1993; Weinert and Schneider 1999; Wellman 1977; Yussen and Bird 1979). To assess early metacognitive knowledge, young children were usually engaged individually in interviews. The children were presented with a series of tasks, and they usually were requested to answer in an open format or else choose one out of two response options. Additionally, justifications for their answers were required from the children (e.g., Ebert 2014; Lockl and Schneider 2006, 2007).

Other investigations addressed the metacognitive understanding of children at the end of elementary school or during their secondary school years (Artelt et al. 2001; Baker and Brown 1984; Schlagmüller et al. 2001; van Kraayenoord and Schneider 1999). For older children, more economic group tests were generally administered. They consisted of various tasks, which were presented verbally in test booklets. So far, few studies have concentrated on metacognitive knowledge in children at early elementary school (Annevirta and Vauras 2001; Fritz et al. 2010) and researchers have not yet examined whether components of metacognitive knowledge, such as different mental processes or strategies, actually build empirically distinguishable subdimensions. The present study, hence, aimed to bridge the gap between research on early declarative metacognitive knowledge in preschool and studies on more elaborative metacognitive concepts at later elementary school. For this purpose, we were at first concerned with the issue of how to appropriately measure children's metacognitive knowledge in early elementary school. Secondly, we were interested in the dimensionality of children's knowledge about mental processes and strategies in the first school years. In the following, we begin with reporting on challenges and difficulties in assessing young children's metamnemonic knowledge. Next, we outline several findings that provide a first insight into the structure of metacognitive knowledge in elementary school children.

Assessment of metacognitive knowledge in early elementary school children

One of the first instruments to assess children's knowledge about memory has been developed by Kreutzer et al. (1975). They used an interview battery comprising 14 items to explore and describe children's knowledge about everyday memory phenomena that tapped a broad variety of metamnemonic areas. The interview contained both open ended questions concerning planning for future retrieval and retrieving past objects or events, and a series of questions with two options about variables influencing memory performance. In the latter type of questions, children were presented with pictures or stories and they had to decide which one was easier to learn or remember. Additionally, children were asked why they had chosen the respective alternatives. By testing preschoolers, first, third, and fifth graders, Kreutzer and colleagues delineated a first picture of children's metamemorial knowledge.

Many subsequent studies on metacognitive knowledge focusing on knowledge about memory processes also used verbal questionnaires or interviews and took portions of Kreutzer's battery. Searching for connections between metamemory and memory, Cavanaugh and Borkowski (1980) made a detailed analysis of the tasks in Kreutzer's battery. They examined correlations between the subtests and memory performance in strategic tasks and looked at developmental differences for each metamemory subtest. Some subtasks showed even negative associations with memory performance, for instance *retrieval*: *event* assessing systematic memory search. In contrast, tasks concerning planned behavior in preparing for future retrieval (*preparation*: *object*; *preparation*: *event*) and tasks assessing knowledge of the effect of increased *study time* were found to be both sensitive and consistently related to memory performance. Furthermore, children's knowledge about organizational strategies for learning verbal material (*study plan*), their recognition that ease of association of item pairs affects learning (*opposites-arbitrary*), and knowledge that gist recall is usually easier than rote recall (*rote-paraphrase*) were sensitive and valid indicators for memory performance.

In the following, researchers were concerned with refining items, extending the item pool and establishing appropriate scoring systems. Borkowski et al. (1983) slightly changed the scoring system for open ended questions, Kurtz and Borkowski (1984) enriched their test with newly developed tasks and Schneider (1986) constructed new items focusing on knowledge about organizational strategies. Most researchers administered different item formats in interviews. Children were required to choose among several response options, to explicate why they had chosen a certain option or to generate useful strategies for learning situations. Investigators also contributed to the assessment of metacognitive knowledge by developing group tests that facilitated the measurement of children's knowledge about mental processes (Belmont and Borkowski 1988; Hasselhorn 1994). Furthermore, tasks that tapped children's memory monitoring and regulating, for instance, predicting their memory span, were included in some test instruments (e.g., Levin et al. 1977).

However, these early tests to assess children's knowledge about memory functioning have several limitations. First, the instruments exhibited rather unsatisfactory internal consistency (Hasselhorn 1994; Kurtz et al. 1982). Moreover, the tests relied on fuzzy conceptualizations, as the instruments constructed by Kurtz et al. (1982) and Belmont and Borkowski (1988), adopted by Hasselhorn (1994), consisted of tasks that assessed children's metacognitive knowledge and items testing their regulatory abilities. Composite scores were formed across all tasks. Yet, there is consensus among researchers now that the knowledge and the regulatory component of metacognition form two distinct domains of metacognition (Fritz et al. 2010; Joyner and Kurtz-Costes 1997), and, therefore, may not be summarized to a single metacognitive knowledge score. After all, the different types of items may partially have measured different abilities of the children. Fritz et al. (2010) reported the occurrence of two

metacognitive knowledge factors, one based on open-ended tasks and the other based on multiple choice (MC) items pointing to different underlying processes. Furthermore, Joyner and Kurtz-Costes (1997) argued that tasks asking respondents to freely generate strategies or give reasons for a certain choice require elaborative productive language skills on behalf of the children and may, thus, cause confounding with verbal abilities and potentially also with personal traits, such as extraversion or openness.

Recently developed test instruments tried to overcome these limitations of early assessments of metacognitive knowledge. To reduce receptive verbal demands of metacognitive tasks, newer tests on the one hand enhanced respondents' understanding of the respective items by giving pictorial presentations of the situations described (e.g., Annevirta and Vauras 2001; Fritz et al. 2010; Lockl and Schneider 2006, 2007). Fritz et al. (2010) individually assessed children's metamemorial knowledge with adapted versions of six declarative metamemory subtests from Kreutzer et al. (1975). Their results showed that illustrations accompanying the tasks helped to decrease language demands on the younger children. On the other hand, pair comparisons and ranking methods instead of open-ended questions were applied to decrease expressive linguistic requirements (Annevirta and Vauras 2001; Lockl and Schneider 2006, 2007; Schlagmüller et al. 2001). Annevirta and Vauras (2001) and Schlagmüller et al. (2001) also enriched the item pool and included tasks assessing knowledge about effective learning procedures and tasks tapping children's knowledge about text comprehension strategies. Moreover, according to the task specific strategy assessment of Schneider (1986), Schlagmüller et al. (2001) implemented new tasks on knowledge about semantic categorization.

Ranking formats, such as the ambitious ranking tasks in the Würzburg metamemory test by Schlagmüller et al. (2001), which has been developed for children at the end of elementary school, may indeed overwhelm young children's memory capacity. Hence, Schneider (1989) suggested providing pair comparisons instead of multiple ranking options for younger children. Regarding the psychometric properties of the newly developed group instrument of Annevirta and Vauras (2001), there were only few discriminating items for first and second graders. Accordingly, no appropriate instrument was available for early elementary school children that could be administered in group settings.

To combine the accomplishments made so far with an economic and suitable assessment in measuring metacognitive knowledge of young elementary school children, a new test was developed within the longitudinal BiKS-3-10 study on Educational Processes, Competence Development, and Selection Decisions at Preschool and Elementary School Age (German: Bildungsprozesse, Kompetenzentwicklung und Selektionsentscheidungen im Vor- und Grundschulalter; von Maurice et al. 2007).[1] The aim was to construct a fair instrument for different subgroups, to cover a wide range of difficulties, and to assess a broad concept of metacognitive knowledge. Items were selected that had been found to be valid and sensitive indicators for young children's metacognitive knowledge. On the basis of analyses by Cavanaugh and Borkowski (1980), Fritz et al. (2010), and Lockl and Schneider (2006), the sensitive and valid items named *preparation*: *object*, *preparation*: *event*, *study time*, *study plan*, *opposites-arbitrary*, and *rote paraphrase* were implemented in the test. According to the results of Schneider (1986), Schneider and Bjorklund (1998), and Schlagmüller et al. (2001), further tasks on semantic categorization were included. Finally, examining the results of tasks on learning and comprehension strategies in the studies of Myers and Paris (1978), Paris et al. (1984), Annevirta and Vauras (2001) and Schlagmüller et al. (2001), further items tapping

[1] The test was developed under the direction of K. Lockl and S. Ebert as part of a subproject (headed by S. Weinert) within the interdisciplinary research group BiKS.

knowledge about ways to learn and understand things were adopted from these studies or newly developed. In contrast to other assessments for this particular age group, children were not tested individually, but in a group setting to facilitate the assessment of young children's metacognitive knowledge and to provide a valuable and economic instrument for classroom teachers. Considering children's low reading competence when starting elementary school, a trained test examiner read each of the tasks to the children. Additionally, the children were presented with illustrations of each of the alternatives in the test booklet. As Fritz et al. (2010) and Saß et al. (2012) had shown, illustrations of tasks helped to reduce language load and time needed to respond to test items. Consequently, the verbal and pictorial information in the newly developed instrument should enhance children's attention and memory by helping them to identify the key features of the problem presented in a task (see also Larkin and Simon 1987). Finally, in order to diminish influences of expressive linguistic skills, no justifications for answers or free generation of strategies were required from the young test takers.

While we wanted to diminish language and memory capacity demands on processing the tasks, we were aware of the theoretical meaningful relation between metacognitive knowledge, language, and intelligence. On the one hand, language abilities play an important role for acquiring metacognitive knowledge and, thus, are inherently related to it (Ebert 2014; Lockl and Schneider 2007). On the other hand, the development of metacognitive knowledge is accompanied by changes in intellectual ability with a constant impact of intelligence on metacognitive knowledge throughout the school years (Alexander et al. 1995, 2006; Swanson 1992). But, although there is an intertwined relation of cognitive and metacognitive processes, several authors outline that metacognition can clearly be distinguished from intellectual ability (Sternberg 1990; Veenman et al. 2006). In the present study, we were interested in empirical results concerning the relations between metacognitive, cognitive, and verbal abilities as part of validating the test.

In conclusion, one goal within our study was to evaluate the newly constructed test of the BiKS study in order to examine whether the ambition of establishing an appropriate and economic instrument for early elementary school children was achieved.

Structure of metacognitive knowledge in elementary school

Recently, Fritz et al. (2010) have empirically confirmed the theoretical division of metacognition into a procedural and a declarative component, which many authors had declared theoretically before (e.g., Brown et al. 1983; Schneider and Lockl 2008; Schneider and Pressley 1997). By conducting factor analyses, they revealed that the regulatory and the knowledge component form two distinct dimensions of metacognition in school-aged children and, thus, recommended against forming a single composite score. Whereas researchers have by now widely agreed on the distinction between these two aspects, information on the structure of the knowledge component itself is still rare.

When we reviewed studies on the structure of metacognitive knowledge in children of different ages, we found few analyses for elementary school children. In studies focusing on secondary school children, metacognitive knowledge is often measured domain-specific reflecting the fact that the development of different competencies, for instance, in English, mathematics, and reading, is paralleled by a certain specialization of the respective metacognitive knowledge (Artelt et al. 2009; Lingel et al. 2010; Schlagmüller and Schneider 2007). To empirically justify the domain-specific measures in secondary school, Neuenhaus et al. (2011) examined the dimensionality of metacognitive knowledge for the domains reading and mathematics in fifth graders. Their results provided evidence for multidimensionality. The two

dimensions of metamnemonic knowledge built distinct components with latent correlations of r=.50 between metacognitive knowledge in reading and math.

Regarding elementary school children, tests on metacognitive knowledge for this particular age group usually consist of several subdimensions and researchers proposed different classification systems to assess metacognitive knowledge in elementary school (e.g., Annevirta and Vauras 2001; Schlagmüller et al. 2001). In their test to individually measure metacognitive knowledge in early elementary school, Annevirta and Vauras (2001) subdivided metacognitive knowledge into three mental processes: they included tasks in which children were asked about the best way to *remember* something, tasks that tapped children's understanding about ways to *learn* new information, and tasks that referred to children's knowledge about the best option to *understand* a certain problem. Schlagmüller et al. (2001) introduced a different classification scheme in their test for children at the end of elementary school based on different strategies. In their instrument on metacognitive knowledge, they differentiated between *general declarative metamemory*, which included knowledge about everyday memory activities and general aspects influencing memory, knowledge about *semantic categorization strategies*, and knowledge about *text processing memory*. Altogether, Annevirta and Vauras' dimensions referred to knowledge about different mental processes, whereas Schlagmüller and colleagues differentiated between different types of strategies. Yet, what has not been analyzed is whether these subdimensions in the tests by Schlagmüller et al. (2001) and Annevirta and Vauras (2001) would empirically form a multidimensional structure of metacognitive knowledge. Instead, they formed composite scores assuming a unidimensional structure. Nevertheless, there are arguments that suggest reconsidering the assumption of unidimensionality of metacognitive knowledge for the respective age period.

Schlagmüller et al. (2001) found only low to moderate intercorrelations between their subdimensions ranging from r=.15 to r=.31. These correlations point to some multidimensionality according to different strategies. At the same time, indicating multidimensionality, partially diverse development of components within metacognitive knowledge was found (Weinert and Schneider 1999). Whereas knowledge about prospective and retrospective retrieval strategies increased substantially in preschool and during the early elementary school years (Lockl and Schneider 2006; Kreutzer et al. 1975), only secondary school children or children at the end of elementary school knew about the usefulness of organizational strategies or facilitations in learning due to relations among paired associations (Flavell et al. 2002; Justice 1985; Sodian et al. 1986). Given these findings, the dimensionality of metamnemonic knowledge still remains unclear. It has not been investigated yet whether metacognitive knowledge as it is assessed in these instruments empirically consists of distinct components such as knowledge related to different mental processes or strategies or whether the construct is rather unidimensional in early elementary school.

Research questions

The first goal of the study was to explore the quality of the new test on metacognitive knowledge developed in the BiKS-3-10 study. Because of the thorough application of a series of criteria in developing the instrument, we suggested that the test would exhibit good psychometric properties and that most of the items would be fair for various subgroups. As has become evident in several studies, language skills as well as cognitive abilities contribute to metacognitive development and are related to it (Alexander et al. 2006; Ebert 2014; Lockl and Schneider 2007). However, due to the decrease in linguistic demands within the tasks, we hypothesized that the relation between metacognitive knowledge and language competencies should be moderate, reflecting that the

actual test performance did not depend heavily on verbal abilities. Furthermore, we expected that tests of metacognitive knowledge and cognitive abilities measured distinct constructs and were not highly associated with each other, as the present test aimed to reduce the requirements on children's memory capacity by using accompanying pictures.

The second goal of the study was to address the issue of dimensionality of metacognitive knowledge in elementary school. Even if researchers subdivided metacognitive knowledge according to mental processes or specific strategies, it has not been explored yet whether these components form empirically distinguishable subdimensions of metacognitive knowledge or whether metacognitive knowledge can be conceived as a unidimensional construct in early elementary school. However, analyses on the structure of a construct are crucial (American Educational Research Association AERA et al. 1999) before addressing research questions concerning the relationship with other variables. Hence, we wanted to examine whether metacognitive knowledge is multidimensional distinguishing between different mental processes (see Annevirta and Vauras 2001) or strategies (see Schlagmüller et al. 2001) or whether it can be considered as a unidimensional construct.

Method

Participants

The present study was carried out with data of the longitudinal project BiKS-3-10. The general focus of the BiKS study is on Educational Processes, Competence Development, and Selection Decisions at Pre-and Elementary School Age. Within a longitudinal design from preschool to elementary school, children in the BiKS-3-10 study participated in a variety of tests. The present study focuses on measurement occasions at the end of first grade and one year later at the end of second grade. In total, the sample consisted of 886 children in first grade and data of 740 children were available in second grade. For the analyses, 16 children in first grade and 20 children in second grade with less than three valid responses in the metacognitive knowledge test were excluded from the analyses since no reliable competence score could be estimated for them. Thus, 870 children in first grade (53.5 % girls) and 720 children (52.4 % girls) in second grade were included in the analyses. Overall, 664 of them participated in the test session in the first as well as in the second grade. The most common reasons for failing to participate in the study in the second grade were that parents had lost interest in the study or that the family had moved away so that their children attended another school. Children were about 7 and a half years old (M= 89.1 months, SD=4.6 months) at the first measurement occasion and 8 and a half years old (M= 99.4 months, SD=4.4 months) at the second measurement occasion. For 25 children, information about age was not available at the second measurement point. The longitudinal BiKS project had already started in preschool. Additional to children who had already been tested in preschool new children were included in the study at transition to elementary school. Specifically, 385 of the first graders in the present study had already been assessed in preschool, and 485 first graders participated for the first time. All children had written parental consent for participation in the study. They attended schools in the two federal states Bavaria and Hesse. While 77.4 % (76.9 %) of the children in first (second) grade had a German language background (no migration status), 19.5 % (20.7 %) of the participants themselves or at least one of their parents spoke another native language than German. For 3.1 % (2.4 %) of the test takers, missing values occurred on migration variables. In terms of socioeconomic status, the highest value of the International Socioeconomic Index of Occupational Status (ISEI; see Ganzeboom et al. 1992) in each family varied between 16 and 90 (M_{t1}=53.33, SD_{t1}=15.75; M_{t2}=53.13, SD_{t2}=15.98). 10.8 % (10.0 %) children had a

missing value on this variable. With respect to the mother's education, 19.5 % (20.3 %) of the mothers had no degree or degree at vocational level in first (second) grade, 35.4 % (35.1 %) had a general certificate of secondary education, and 34.3 % (33.2 %) had qualifications for university entrance. For 6.4 % (6.3 %) of the children information about the mother's education was not available.

Materials and procedure

At each measurement point, children participated in the testing in a group setting with a group size of about ten children. A trained examiner administered tests measuring verbal abilities, nonverbal cognitive abilities and metacognitive knowledge. To ensure that the children understood the instructions of the various instruments, sample items were used for each test and the experimenter explained the respective procedures. Children who were absent at the time of testing and who had already participated longitudinally in preschool were visited at home and tested individually. In both settings, the tests and the items in the tests were presented to all children in the same order. The present study focused on four measures: the newly developed instrument to assess children's metacognitive knowledge and three measures to evaluate discriminant validity, that is, a cognitive nonverbal ability test (CFT 1; Cattell and Osterland 1997), a sentence comprehension test assessing grammar (TROG-D; Fox and Bäumer 2006), and a receptive vocabulary test (subtest of KFT 1–3; Heller and Geisler 1983).

Metacognitive knowledge

The children received the same test for metacognitive knowledge in first and second grade. A pilot study with a larger item pool had been carried out beforehand and 15 items with good psychometric properties covering a broad range of difficulties were selected for the final test instrument. Testing lasted about 15 min and the test consisted of 15 MC items, which were partially taken or adapted from other studies (Kreutzer et al. 1975; Lockl and Schneider 2006; Wellman 1977; Schlagmüller et al. 2001). For each of the 15 tasks on metacognitive knowledge, a situation involving mental performance and three options were presented to the children. The test examiner read aloud the situations and the corresponding options and the children followed each approach by looking at the pictures in their test booklet. The examiner then asked the children which of the options presented they thought would be the best for performing a particular task. The children had to mark one out of the three options. Two of the options always showed two different ways of acting in the given situation or different conditions for mental performance. Children also had the possibility to choose the third option stating that the two presented alternatives work equally well. For each item, there was one option being the best with reference to the items of the previous studies mentioned above. Either one of the alternatives with differing strategic quality was better to act in the given scenario or the two alternatives were equally good. Children were rewarded with one point, if they chose the correct answer, otherwise, they got zero points. An example of a test item is depicted in Fig. 1. As Schneider (1989) recommended, this type of pair comparison was used instead of ranking several options so as not to overburden the young children's memory capacity.

A variety of tasks has been incorporated in the new metacognitive knowledge test. All items are given in a shortened, not verbatim, version in the Appendix. Items were chosen to appropriately reflect the various aspects of metacognitive knowledge that experts in that field had been considered as important in previous research (Annevirta and Vauras 2001; Cavanaugh and Borkowski 1980; Kreutzer et al. 1975; Schlagmüller et al. 2001; Schneider

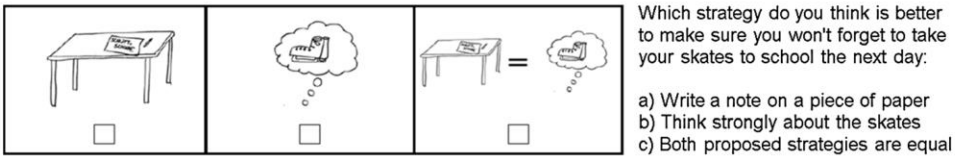


Fig. 1 Example for a MC item in the metacognitive knowledge test

and Bjorklund 1998). The items can be assigned to different classification schemes. On the one hand, the tasks assessed metacognitive knowledge about the mental processes *remembering*, *understanding*, and *learning*. This classification scheme has been established by Annevirta and Vauras (2001) and points to the cognitive functions involved in memory activities. According to Annevirta and Vauras' classification the newly developed test consisted of five items to capture children's knowledge about factors affecting *remembering* in various situations (Items 1, 2, 6, 8, 10), five items to gather knowledge about ways to maximize *understanding* of texts or tasks (Items 3, 7, 11, 14, 15) and, finally, five items to assess knowledge about effective *learning* procedures (Items 4, 5, 9, 12, 13). Besides, the tasks can also be subdivided into knowledge about different types of strategies (according to Schlagmüller et al. 2001). Tasks on knowledge about strategies were included in the test, which were the most sensitive and valid indicators for children at early elementary school (Kreutzer et al. 1975; Cavanaugh and Borkowski 1980; Schlagmüller et al. 2001). Accordingly, the items in the test can be attributed to *everyday mental activities* (Items 1, 2, 4, 6, 11), items referring to *semantic categorization strategies* (Items 7, 9, 10, 12, 13), and items testing *school related metacognitive knowledge* (Items 3, 5, 8, 14, 15).

The items of the dimension *everyday mental activities* (EMA) referred to knowledge about prospective or retrospective retrieval strategies that persons might use to remember information, and to everyday personal knowledge, such as the impact of previous knowledge on performing in a task. To measure children's knowledge about retrieval strategies, the children were, for example, given two alternatives for remembering to take their skates to school the next day (see Fig. 1). Children were then asked which one of the strategies they found was better or whether they thought that the strategies were equally good.

The items belonging to the *semantic categorization strategies* (SCS) required knowledge about the advantages of organizing learning material into meaningful clusters. For example, participants were given two lists of identical pictures to remember. In one list the pictures were presented in two conceptual categories, arranged as clothes and toys, whereas the other list contained the items in randomized order.

Finally, items that assessed *school related metacognitive knowledge* (SRM) were included in the test. The items required knowledge about the impact of study time on learning or about rehearsal and text comprehension strategies.

General cognitive nonverbal abilities

To assess children's cognitive nonverbal abilities, the Culture Fair Test (CFT 1; Cattell and Osterland 1997) was administered. The short form embedding the subtests classifications, similarities, and matrices was used in the testing. As the test manual suggested, a composite score was built on the basis of the three subtests resulting in a maximum score of 36.

Language abilities

Sentence comprehension (grammar) was assessed with the German version of the Test for Reception of Grammar (TROG-D; Fox and Bäumer 2006). The short version of the test used in this study consisted of 44 items, and a maximum of 44 points could be reached. In addition, the subtest to assess verbal abilities of the Cognitive Abilities Test (KFT 1–3; Heller and Geisler 1983) was used to measure children's receptive vocabulary (maximum score of 15).

Analyses

To analyze the data, we used different models of Item Response Theory (IRT). Especially in the field of test evaluation, there are beneficial properties and possibilities associated with the application of IRT (see, e.g., Embretson 1996). Specifically, it can be tested whether or not the items fit the underlying model, differential item functioning can be explored, and, in case of a good fit of the model to the data, precise ability estimates with different standard errors at each trait score level can be computed (Embretson and Reise 2000; Thissen and Wainer 2001). Different IRT analyses were performed using the Software ConQuest (Wu et al. 1997) to investigate the psychometric properties of the newly developed test. The Rasch model was chosen to model the data as it preserved the equal item weighting of the 15 items intended by the test developers. Marginal maximum likelihood estimation implemented in ConQuest was applied for estimating item and person parameters. In case of item-non response, this estimation approach ensures including all available data to estimate the model parameters (e.g., Enders and Bandalos 2001; Little and Rubin 1987). Rather few missing values within the test occurred and, as suggested by Pohl et al. (2013), they were ignored in the IRT analyses. The fit of the items to the Rasch model was evaluated based on the weighted mean square (WMNSQ), the respective *t*-value, the point biserial correlation between the item and the total score, and the item characteristic curve. As another crucial criterion of test quality (e.g., American Educational Research Association AERA et al. 1999), test fairness was checked, that is that no items favor certain subgroups e.g., are easier for boys than for girls when controlling for overall group differences on the latent trait. In order to check for test fairness in the newly developed test instrument, differential item functioning (DIF) analyses were conducted and the size of DIF was evaluated.

Within the evaluation of construct validity also discriminant validity of the test was investigated by correlating the metacognitive knowledge score with measures of language skills and cognitive nonverbal abilities.

The structure of the test was examined by applying two three-dimensional models to the metacognitive knowledge data. In terms of structural equation modeling (SEM) the IRT models can be seen as confirmatory factor analyses that test the structure of the construct. The first model consisted of the strategy dimensions EMA, SCS, and SRM. The second model was based on the mental processes *learning*, *remembering*, and *understanding*. In order to compare the fit of the multidimensional models and the unidimensional model, a χ^2-difference test was computed to explore significance of differences between the better fitting multidimensional model and the unidimensional model. Since the sample size is quite large, also small differences in model fit between the two models can get significant. Thus, additionally two widely applied descriptive fit indices were used: the Akaike's (1974) information criterion (AIC) and the Bayesian information criterion (BIC, Schwarz 1978). They are both model fit measures with the BIC penalizing additional parameters in the model (complex models) more than the AIC. The smaller the value of AIC or BIC, the better is the model.

We investigated the homogeneity of change in the underlying construct as further indicator for dimensionality. Beforehand, analyses on measurement invariance from first to second grade were undertaken (e.g., Millsap 2010) to ensure that the components' structure has not changed. Therefore, a model with constraint item parameters between first and second grade for the three dimensions EMA, SCS, and SRM was compared to the multidimensional baseline Rasch model without constraints on item parameters. Subsequently, a multidimensional latent change model was applied to the data and latent correlations between the factors of change were inspected. The latent change variables were modeled as difference in the true-score variables of both occasions of measurement (Steyer et al. 1997). The aim of this analysis was to explore whether subjects who improve more on the one dimension are also likely to improve more on other dimensions or whether changes are not highly related (see e.g., Vautier and Pohl 2009, for the latent change approach for examining dimensionality). As Vautier and Pohl (2009) state, change variables that are not perfectly correlated provide evidence for multidimensionality and reflect that the construct encompasses subdimensions that evolve differently.

Results

In the following, we first briefly report a) descriptive statistics of the metacognitive, cognitive, and language performance in first and second grade. We then show b) the results of the evaluation of the newly developed test to measure metacognitive knowledge using different IRT-models. Finally, we present c) the results regarding the structure of metacognitive knowledge.

Descriptive statistics of metacognitive, cognitive, and language performance

Table 1 shows the mean scores of the metacognitive knowledge test, the test to assess general cognitive abilities, and the language instruments to measure children's receptive grammar and vocabulary. As expected, children's metacognitive, cognitive, and language abilities increased substantially between first and second grade. In first grade, children solved about half of the metacognitive tasks correctly, in second grade, they solved almost three quarters of the items, $t(664)=-17.91$, $p<0.01$, $d=0.75$. Significantly higher average scores in second than in first

Table 1 Descriptive statistics of children's competencies across Grades 1 and 2

	Time 1: first grade Mean (*SD*)[a]	Time 2: second grade Mean (*SD*)[a]
Metacognitive knowledge (0–14)[b]	7.72 (3.07)	9.58 (2.60)
General cognitive abilities (0–36)	26.01 (5.45)	27.89 (4.04)
Grammar (0–44)	34.27 (5.35)	37.43 (4.27)
Vocabulary (0–15)	7.09 (2.51)	8.68 (2.61)

[a] Standard Deviations are in parentheses

[b] One of the 15 items in the test was excluded from the analyses due to unsatisfactory item fit (see section "evaluation of the metacognitive knowledge test")

grade were also present for general cognitive abilities, $t(640)=-9.42$, $p<0.01$, $d=0.35$, grammar, $t(635)=-22.49$, $p<0.01$, $d=0.68$, and vocabulary, $t(638)=-18.302$, $p<0.01$, $d=0.62$.

The descriptive statistics of metacognitive knowledge give a first impression of the tests' difficulty and the development of children's knowledge about memory processes in first and second grade. In the following, the IRT analyses provide information at item level about the quality of the metacognitive knowledge test.

Evaluation of the metacognitive knowledge test

In order to examine the psychometric properties of the newly constructed test, first a Rasch model was applied to the data and item fit statistics were explored. Detailed results of the Rasch analyses for each item including item difficulty, item fit, and DIF, are depicted in the Appendix. Regarding the amount of correct answers for each item (an indicator for the item's difficulty), there was a considerable variation in the probability of solving the items in first (second) grade ranging from 33.9 % (45.1 %) to 75.2 % (84.7 %), $M_{t1}=56.2$ %, $M_{t2}=68.3$ %. For the first grade, the items particularly exhibited a medium difficulty and, thus, metacognitive knowledge of most of the persons was measured very precisely. In the second grade, the probability of solving the items increased and especially for subjects with a low and medium ability precise ability estimates were obtained. Overall, items referring to EMA (Items 1, 2, 6, 11) were easier for children than items from the other dimensions. Items concerning SCS (Items 7, 9, 10, 12, 13) had a medium probability to be solved. In contrast, tasks assessing SRM, which includes for example tasks to measure text comprehension knowledge and knowledge about rehearsal strategies (Items 5, 14, 15) were the most difficult in both grades.

To evaluate the fit of the items to the model, we checked the WMNSQ with the respective t-value, point biserial correlations of the items with the total score and the item characteristic curves. The WMNSQ of an item describes the deviation of the observed and the model implied probability for a correct response given a certain ability level (Wright and Masters 1982). A WMNSQ close to 1 indicates that the item fits the measurement model well. Similar to rules of thumbs given in other large scale studies (e.g., Adams and Wu 2002; Martin et al. 2004; Pohl and Carstensen 2012) taking the sample size into account, an item fit of $0.90<\text{WMNSQ}<1.10$ and a point biserial correlation >0.30 was considered as good. 14 of the 15 items exhibited a very good item fit with WMNSQ ranging from 0.93 to 1.05, point biserial correlations varying between 0.29 and 0.55, and a good fit of the empirical item characteristic curves to the model implied curves. Item 4 (*Irrelevant*) showed rather low point biserial correlation ($r_{t1}=0.14$, $r_{t2}=0.22$), an unsatisfactory fit ($\text{WMNSQ}_{t1}=1.22$, $\text{WMNSQ}_{t2}=1.13$), and the empirical item characteristic curve deviated considerably from the model implied curve in first as well as in the second grade. This item was, therefore, excluded from further analyses.

The EAP/PV reliability of the metacognitive knowledge test based on 14 items was acceptable with .68 and .62 at measurement point 1 and measurement point 2, respectively. THE EAP/PV reliability is an estimate for IRT test reliability obtained by dividing the variance of the expected a posteriori ability estimates by the estimated total variance of the latent ability (ratio of modeled variance to observed variance).

Fairness of the test was checked by conducting DIF analyses for the variables gender, migration, and design (i.e., whether children had already participated in the BiKS study in preschool or were newly included when starting elementary school). The size of DIF was evaluated with respect to the classification scheme given by the Educational Testing Service (Zwick et al. 1999), and items were judged as having a small, moderate, or large DIF. For almost all items rather small DIF below 0.43 logits emerged. Six items exhibited a moderate

DIF (DIF>0.43 logits, significantly deviating from zero). However, in neither grade there was an item with large DIF greater than 0.64. Only one item had a DIF of nearly 0.64 logits, namely Item 14 (*Story I*) in first grade. This item was about 0.6 logits more difficult for children without migration background than for children with migration background who had the same overall ability. Indeed, exploration of the item's content yielded no indication of unfairness. The assumption of this DIF being rather accidental was supported by the fact that the DIF value in second grade was far below 0.64 logits. In sum, test fairness for the variables gender, migration background, and design was confirmed as a large number of items showed negligible DIF and no item had a strong DIF.

Finally, discriminant validity of the test was studied. For this purpose, the metacognitive knowledge score was correlated with relevant external variables (see Table 2).

Moderate correlations between metacognitive knowledge and language skills occurred, supporting the hypothesis that the two tests measured different constructs. The correlations showed a substantial relation between declarative metacognition and verbal abilities with higher correlations for sentence comprehension (grammar) than children's vocabulary. Because cognitive abilities are substantially related to both metacognitive knowledge as well as language skills, we also investigated the relationship between metacognitive knowledge and verbal competencies, thereby controlling for general cognitive nonverbal abilities. Both partial correlations remained significant. Altogether, the present correlations exhibited a significant association between linguistic competencies and metacognitive knowledge, but their value indicated that the test on metamnemonic knowledge measured a construct that is empirically well distinguishable from tests assessing language skills.

Looking at the relation of metacognitive knowledge and general cognitive abilities, moderate correlations could be found, which were in line with the assumption that the test did not measure cognitive skills in particular. Although substantial associations were present in first and second grade, the degree to which the two abilities were related clearly suggested that the test assessing nonverbal cognitive abilities and the test on metacognitive knowledge captured distinct constructs. Overall, the analyses showed a good discriminant validity of the test and strengthened the psychometric quality of the new instrument.

All items except for Item 4 (*Irrelevant*) in both grades were finally retained in the instrument and subsequent analyses were, thus, based on 14 items. Note that due to the deletion of Item 4, the dimensions *knowledge about everyday mental activities* and the mental process of *learning* consisted of only four instead of five items.

Structure of metacognitive knowledge

In order to examine the structure of children's metacognitive knowledge at the beginning of elementary school, we specified two three-dimensional models: one model based on different

Table 2 Discriminant validity of the metacognitive knowledge test

Correlation of metacognitive knowledge with…	first grade	second grade
Language skills		
Grammar	.29** (.17**)[a]	.32** (.16**)[a]
Vocabulary	.18** (.11**)[a]	.17** (.10**)[a]
General cognitive abilities	.29**	.26**

[a]Correlations in parentheses are partial correlations controlling for general cognitive abilities
*p<.05, **p<.01

strategies (adapted from Schlagmüller et al. 2001) and the other reflecting mental processes (Annevirta and Vauras 2001). Table 3 presents the overall fit indices AIC and BIC of the two three-dimensional models and the unidimensional model.

In the first two columns, AIC and BIC of the first grade are depicted. In the second two columns, AIC and BIC of the second grade are shown. For comparing the fit of the models, the rows in one column have to be looked at. At first, we regarded the two three-dimensional models. In first as well as in second grade AIC as well as BIC had lower values for the model that distinguished between different types of strategies than for the model that subdivided into different mental processes. This indicated that the empirical data were modeled better by the classification scheme differentiating between strategies.

Next, the fit indices of the better fitting three-dimensional model based on strategies and the unidimensional model were compared. For the first grade, AIC preferred the multidimensional model based on strategies, BIC values were slightly lower for the unidimensional model. The differences in the deviances yielded statistical significance, $\chi^2(5, N=870)=33.38$, $p<.01$ indicating that the multidimensional model fitted the data significantly better. For the second grade, AIC and BIC favored the multidimensional model and the change in deviance again attained statistical significance, $\chi^2(5, N=720)=38.24$, $p<.01$. Taken together, almost all fit criteria indicated a better fit of the three-dimensional model except for the BIC in first grade.

In addition to overall fit indices, latent correlations, that is intercorrelations corrected for measurement error, among the strategic dimensions were taken into account (see Table 4) as important indicator for dimensionality. Overall, slightly higher correlations between the dimensions occurred in first grade than in second grade. The components that showed the highest correlations differed from first to second grade, whereas in both grades the lowest correlations were found among SCS and EMA. Three out of six correlations were lower than 0.85, which again points to multidimensionality of the construct and suggests that the variables are not likely to measure the same latent trait (see Kline 2005).

After examining dimensionality separately for first and second grade, we analyzed the uniformity of the metacognitive development across first and second grade as another indicator for dimensionality. Therefore, we first regarded the development of the strategic subdimensions and we then explored the homogeneity of the development. Before applying latent change models to the data in order to evaluate the homogeneity of the component's development, tests to assess measurement invariance of the strategic dimensions from first to second grade were conducted. At least strong measurement invariance, that is equality of item parameters (Widaman and Reise 1997), must hold to justify that the same latent construct is measured at different measurement occasions (Widaman et al. 2010). Except for three items, measurement invariance constraints were satisfied and supported that the items of the three dimensions EMA, SCS, and SRM reflected the same constructs at both grades. For three items,

Table 3 Fit indices of the multidimensional models and the unidimensional model

Model	first grade		second grade	
	BIC	AIC	BIC	AIC
Unidimensional model	14707.4	14635.8	11269.6	11200.9
Model based on strategies	14707.8	14612.5	11264.3	11172.7
Model based on mental processes	14738.5	14643.1	11298.5	11206.9

Table 4 Latent correlations of the multidimensional model based on strategies for first grade (below diagonal) and for second grade (above diagonal)

	EMA	SCS	SRM
EMA (N_{items}=4)	–	0.75	0.89
SCS (N_{items}=5)	0.80	–	0.85
SRM (N_{items}=5)	0.87	0.94	–

Analyses based on complete cases for both grades showed only very slight deviations of the present correlations ranging from 0.001 to 0.021

EMA everyday mental activities, *SCS* semantic categorization strategies, *SRM* school related metacognitive knowledge

the item difficulty changed across measurement occasions indicating that the item functions in a different way across waves. Thus, further dimensionality analyses were undertaken with the 14 items in both grades applying a model with partial strong invariance (McArdle and Cattell 1994) in which equality constraints between grades were only posed for 11 items. This relaxes the assumption that all items need to measure the same construct across time and the link between the grades is only based on 11 items. All components showed a substantial development from first to second grade with EMA, $t(653)=-.559$, $p<0.01$, $d=0.53$, SCS, $t(653)=-.800$, $p<0.01$, $d=0.90$, SRM, $t(653)=-.437$, $p<0.01$, $d=0.39$ with strongest improvements appearing within SCS. At last, a multidimensional latent change model was specified in order to inspect the latent correlations among the change factors (see Fig. 2). If changes on the latent components of the metacognitive knowledge test were rather low correlated, further evidence for multidimensionality would be given displaying that there is no homogeneous development on all subdimensions.

The high correlations between the change factors, that is, between dimensions $EMA_{t2\text{-}t1}$, $SCS_{t2\text{-}t1}$, and $SRM_{t2\text{-}t1}$ gave evidence for a rather high homogeneity regarding change in metacognitive knowledge. In particular, high correlations between the change of SRM and SCS were obtained ($r=.91$). Subjects whose competence on SCS increased tended to show higher change in SRM, too. Lower correlations, underlining certain multidimensionality, were present between SCS and EMA ($r=.85$) and the correlation between change in EMA and SRM was $r=.89$. Overall, the more children acquired knowledge on one dimension the more they improved on other dimensions,

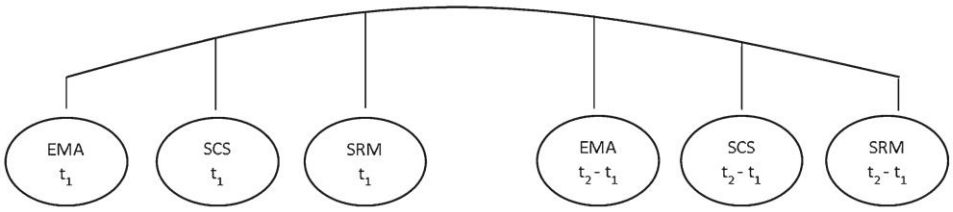


Fig. 2 Multidimensional latent change model. *EMA* Everyday mental activities, *SCS* Semantic categorization strategies, *SRM* School related metacognitive knowledge. Note, that the measurement models of the latent variables are not displayed in the figure

too. The overall fit indices presented unequivocal results: the model assuming a multidimensional change (AIC=25644.30, BIC=25842.36) was preferred over the model with a unidimensional change (AIC=25660.09, BIC=25795.35) by the AIC, whereas the BIC favored the more parsimonious unidimensional change model. The chi-square deviance test was statistically significant at the 0.05 significance level, $\chi^2(28, N=926)=41.79$, $p=.045$. The fact, that the multidimensional change model was not clearly preferred and the high correlations between the change factors indicated that the change on the dimensions within the children was rather homogenous. This supports the assumption of a unidimensional construct.

Discussion

The objective of the present study has been to broaden knowledge about metacognitive knowledge in children at early elementary school as previous studies assessing metacognitive knowledge have predominantly focused on preschoolers or children at the end of elementary school or at secondary school. First, we evaluated a new test for group settings which assessed young children's knowledge about mental activities. Furthermore, to find out more about the nature of metacognitive knowledge for the respective age group, structure and development of metacognitive knowledge were studied across first and second grade of elementary school.

Evaluation of the metacognitive knowledge test

First of all, the new instrument on metacognitive knowledge has beneficial economic features since it can be easily administered in group settings and takes only 15 min. In addition, the test exhibited good psychometric properties including very good item fit for 14 of the 15 items and acceptable reliability of the measures. The final instrument, thus, comprised 14 items assessing a wide range of metacognitive knowledge from leisure time and school contexts. The tasks that tapped a wide range of metacognitive knowledge were well targeted toward the specific population, as can be seen in the test targeting, especially in first grade. Thus, they yielded differentiated measures of the subjects. Moreover, analyses indicated that measurement invariance was present for the relevant subgroups based on gender, design, or migration background. No contents of items favored certain subgroups when controlling for overall group differences and, therefore, test fairness as an important test criterion was achieved. Looking at the strategic subdimensions, we found that children performed substantially better in second grade than in first grade on all dimensions. Thus, children seem to acquire knowledge in all components but do not show huge gains in any of these dimensions in particular. In first as well as in second grade items referring to EMA (e.g., choosing the best option for not forgetting to take the school bag to school the next day) had lower item difficulties than items on SCS with less striking task features. Moreover, tasks on SRM (e.g., knowing the benefits of rehearsal or text processing strategies) had rather high item difficulties in our study. Indeed, item difficulty depends on the construction of the tasks and therefore, difficulties have to be interpreted cautiously. However, the findings are in accordance with previous studies that demonstrated that basic concepts about memory and everyday mental knowledge are already known in early childhood (Annevirta and Vauras 2001; Kreutzer et al. 1975), whereas knowledge about the usefulness of organizational

strategies and text comprehension strategies develops somewhat later (Artelt et al. 2001; Flavell et al. 2002; Justice 1985; Veenman et al. 2006).

With regard to the discriminant validity of the test, the results suggest that neither language skills nor cognitive abilities strongly confounded the measure of metacognitive knowledge. The fact that metacognitive knowledge and language skills were not highly correlated provides evidence of appropriately low verbal demands in the metacognitive tasks. Concerning general cognitive abilities moderate correlations between metacognitive and cognitive competencies occurred suggesting that the newly developed test did also not measure cognitive abilities in particular. At the same time, the correlations between metacognitive knowledge, language, and intelligence correspond to those found in the literature (Alexander et al. 1995, 2006; Grammer et al. 2011; Lockl and Schneider 2006, 2007; Weinert and Schneider 1999). It seems that the variables describe related, but distinct constructs in elementary school children. Taken together, the relations in the present study clearly confirmed the discriminant validity of the test and were consistent with theoretical accounts about the relationship between metacognitive knowledge, linguistic competencies, and cognitive abilities.

Dimensionality of metacognitive knowledge

In order to understand the nature of metacognitive knowledge in early elementary school children, an important aim of the study was to analyze the dimensionality of children's metacognitive knowledge and its development across first and second grade. First, the analyses revealed that the empirical data structure was better represented by the model based on strategies that are used in different contexts than by the model based on mental processes. It seems that knowledge about a certain strategy is more consistent within the children than knowledge about a certain mental process like knowledge about factors influencing *remembering*. Regarding the latent correlations among the strategy dimensions and the significant differences in deviance between the uni- and the multidimensional model, evidence for multidimensionality emerged. The correlations significantly differed from perfect correlations supporting the assumption that children's metacognitive knowledge comprises distinct components. Accordingly, the results are in line with theoretical assumptions of interrelations among different metacognitive knowledge dimensions being weak in young children (Flavell and Wellman 1977; Schneider 1985). The lowest correlations among the dimensions were found between EMA and SCS for both grades. Children who have more knowledge about EMA than others might not necessarily know more about the benefits of SCS in performing a task. Comparing the manifest intercorrelations in the present study for first and second graders with the manifest intercorrelations of Schlagmüller et al. (2001) assessing third and fourth graders, the dimensions of Schlagmüller and colleagues (r=.15–.31) correlated lower than the dimensions in the present study (r=.29–.47). Whether this is due to differences in the test instrument or whether the heterogeneity of children's metacognitive knowledge increases across elementary school, as assumed for the development of metacognitive skills (Veenman and Elshout 1999), is a task for future research. Looking at the results of the latent change model, some correlations among the change factors were rather high. They indicated that children show a similar development on all dimensions. But again, the correlations differed from a perfect correlation, especially among EMA and SCS, underlining the hypothesis of the construct's multidimensionality.

Interestingly, the subdimensions themselves showed a rather high stability in their measurement structure from first to second grade as evident in the analyses on measurement invariance. The results demonstrated that the constructs of the subdimensions did not change considerably across early elementary school allowing for analyzing their trajectories longitudinally.

The finding that metacognitive knowledge in young elementary school children has a multidimensional structure has implications for the new test instrument as well as for further analyses based on measures of metamnemonic achievement. First, the test was evaluated assuming a unidimensional structure based on previous research that had used unidimensional models for evaluating tests on metacognitive knowledge in elementary school (Annevirta and Vauras 2001; Lockl and Schneider 2006; Schlagmüller et al. 2001). Based on a multidimensional structure, the results of evaluation would probably differ slightly and, therefore, future research might investigate the test characteristics with an underlying multidimensional structure. Second, when this test is used in practical settings to assess children's declarative metacognition and when further analyses with metacognitive knowledge data are conducted, thorough considerations on how to adequately model the data should be made. For this purpose, empirical as well as theoretical aspects have to be taken into account when analyzing metacognitive knowledge in young elementary school children. From an empirical point of view, the rather low intercorrelations between dimensions that emerged in the present test and in other tests measuring metacognitive knowledge in elementary school (Schlagmüller et al. 2001) point to the fact that unidimensionality does not fully hold within metacognitive knowledge tests. The use of a unidimensional model for the multidimensional data might, thus, result in a composite that does not distinguish between the different, distinct subdimensions anymore. Consequences of such a model misspecification may be a higher standard error of the person ability estimate and the proficiency classification (Walker and Beretvas 2003). Furthermore, different aspects of metacognitive knowledge may be differently related to other variables and, thus, a multidimensional model can provide a more accurate and detailed picture of these associations. From a theoretical perspective on the other hand, it can be argued that different heterogeneous metacognitive aspects are included in the test in order to assess a wide and comprehensive construct of metacognitive knowledge. A unidimensional model may then be sufficient as approximation for obtaining an overall estimate of the respective competence (Ercikan 2006). With regard to model parsimony in the present study the stricter BIC penalizing additional parameters preferred the unidimensional model in first grade. Based on this stricter value, it might be also comprehensible to choose the unidimensional model instead of a multidimensional model in consideration of the parsimony criterion. Summarizing, there may be arguments that justify forming a single score for metamnemonic knowledge, even if there is empirical evidence that the instrument measures a multidimensional construct. Consequently, depending on the research questions and aims in assessing metacognitive knowledge, both ways of modeling the metacognitive data may be appropriate.

Future research and limitations

With regard to further analyses, promising research questions arise from the analyses on dimensionality. On the one hand, the question of diversity among the strategic factors should be addressed by investigating their relationship with other variables over time. Analyses about factors and competencies that are supposed to influence metacognitive knowledge, such as general cognitive abilities (Alexander et al. 1995;

Bjorklund 2005), verbal abilities (Lockl and Schneider 2007; Schneider et al. 2004), certain home-level factors, for instance, the maternal education level (Grammer et al. 2011), or metacognitive instruction (Pressley and Gaskins 2006; Veenman et al. 1994) might reveal different impact on the metacognitive components. Moreover, it might be of interest to explore specific contributions of the metacognitive dimensions on school performance at early elementary school. The finding that children had a higher knowledge on all dimensions in second grade than in first grade may support the assumption that teachers probably do not explicitly teach specific strategies at school (Pressley et al. 1989). However, to answer this question, detailed analyses on teaching in classes and children's acting in learning environments are necessary.

Limitations of the present work concern the analysis of only two measurement occasions. Longer periods of time would deepen knowledge about change of metamnemonic knowledge. Moreover, we are aware that the analyses on dimensionality depend on the items chosen for the test. Thus, a further limitation concerns aspects of metacognitive knowledge that were not assessed with the metacognitive knowledge test. Although we argue that we tried to incorporate a broad and representative sample of items based on previous research, we could not include all existing scenarios because of time restrictions and decline of attention within the children. Future research with additional scenarios would provide further information about relations between the assessed and other aspects of metacognitive knowledge. Another limitation of the present study is that no other measures of metacognition such as metacognitive monitoring or regulation processes were assessed within the children. The inclusion of measures on procedural aspects of metacognition in subsequent research (using the newly developed metacognitive knowledge test) would allow answering broader questions about relations of the impact of metacognition on other cognitive and school related competencies. This seems especially promising because recent studies show that not only metacognitive knowledge but also metacognitive monitoring and control processes are influential for academic achievement (e.g. Krebs and Roebers 2010; Roebers et al. 2009; Roebers et al. 2012). Furthermore, other measures to establish external validity of the metacognitive knowledge test should be assessed in further research. Previous studies that contained items of the new test instrument such us Cavanaugh and Borkowski (1980), Schneider (1986) and Schneider and Bjorklund (1998) had already shown a meaningful relation between these items and memory performance. Evidence of external validity for the new instrument, for instance, an association between sorting in a task or the application of text comprehension strategies and the metacognitive knowledge test would further strengthen the quality of the instrument. Also for investigating the different subdimensions of metacognitive knowledge more thoroughly, the number of items within the dimensions should be increased in subsequent studies.

Despite these limitations, the present study increases our understanding of the nature of metacognitive knowledge in early elementary school. Hence, the work serves as fundamental step for future research exploring the relationship between metacognitive knowledge or subdimensions of it and other relevant variables.

Acknowledgments The study presented is conducted in the subproject "Analysis of the relationship between acquisition and cognitive development, and acquisition of self-regulative skills and characteristics of adult-child interaction" (Prof. Dr. Sabine Weinert), focusing on educational and psychological research questions. The subproject is part of the larger interdisciplinary research group BiKS, funded by the German Research Foundation. We would like to thank all participating children and their parents as well as all students engaged in data collection for their most active cooperation.

Appendix

Table 5 The 15 items of the metacognitive knowledge test. All items were read aloud to the children by the test examiner. For each item, the children were asked which option might be better for performing a given task or whether they thought both options were equally good

Item number	Label	Item content
Item 1	Birthday party	Which option is better for not forgetting a friend's birthday party? a) Often think about it b) Mark it in the calendar
Item 2	Gym bag	Which option is better for not forgetting to bring the gym bag to school the next day? a) Hang the bag on the door b) Tell my little brother to remind me
Item 3	Sentences	Which sentence is easier to keep in mind? a) The tired girl sleeps all night long. b) The stupid table sings a thick cup.
Item 4	Irrelevant	Which option is better to learn a set of pictures? a) Wearing a stripy t-shirt b) Wearing a t-shirt with polka dots
Item 5	Visit to the zoo	Which option is better to memorize animal names during a visit to the zoo? a) Saying each animal name once b) Repeating the animal names again and again during your stay in the zoo
Item 6	Ice skates	Which option is better for not forgetting to take your ice skates to school the next day? a) Write a note on a piece of paper b) Think strongly about the skates
Item 7	Word pairings	Which word pairings are easier to memorize? a) black-white, big-small, b) wide-yellow, grey-round
Item 8	Study time	Which option is better to remember a set of words? a) Spend 1 min memorizing b) Spend 5 min memorizing
Item 9	Animals I	Which option is better to learn animal names from a book? a) Draw the animals at random b) Draw the animals ordered by category
Item 10	Animals II	Which option is better to remember many animals? a) Remember them by category b) Remember them in random order
Item 11	Familiarity	Who will be better in a quiz about the forest? a) A boy who likes playing with cars b) A boy who likes watching shows about animals and plants
Item 12	List of words	Which option is better to learn a list of words? a) Learn them in randomized order b) Learn them by category
Item 13	Memory cards	Which option is better to memorize memory cards? a) Sort the memory cards b) Randomly repeat them
Item 14	Story I	Which option is better to keep a story in mind? a) Underline the first and last sentence b) Tell the story to your mum and read it again
Item 15	Story II	What is more important for memorizing a story? a) Understand the story b) Read the story fast

Table 6 Item parameters for the metacognitive knowledge test in first and second grade. The Rasch model was estimated by constraining the mean of the latent ability to be zero

Label	Grade	Percent correct	Difficulty[a]	WMNSQ	*t*-value (WMNSQ)	Point biserial correlation	DIF_ gender[ab]	DIF_ design[ab]	DIF_ migration[ab]
Item 1	1st	75.20	−1.258	1.01	0.4	0.37	−0.210	0.082	−0.086
	2nd	82.43	−1.720	1.04	0.6	0.29	−0.330	−0.048	−0.052
Item 2	1st	70.50	−0.991	1.00	0.1	0.42	−0.020	−0.154	−0.120
	2nd	83.54	−1.809	1.00	0.1	0.34	0.042	−0.026	−0.528
Item 3	1st	65.46	−0.728	0.99	−0.2	0.43	0.502**	0.098	0.154
	2nd	72.38	−1.080	1.03	0.8	0.35	0.524**	−0.058	−0.118
Item 4	1st	52.79	−0.128	1.22	8.5	0.14	0.396**	0.160	0.080
	2nd	63.58	−0.626	1.13	3.9	0.22	0.266	0.078	−0.040
Item 5	1st	38.07	0.560	1.05	1.8	0.36	−0.310	0.314	0.250
	2nd	45.12	0.223	1.05	2.1	0.35	−0.308	0.152	0.136
Item 6	1st	69.92	−0.957	0.99	−0.4	0.44	−0.306	−0.068	−0.096
	2nd	81.47	−1.653	0.96	−0.7	0.44	−0.306	−0.244	0.118
Item 7	1st	64.74	−0.696	0.95	−1.6	0.49	0.026	0.100	−0.104
	2nd	81.64	−1.663	0.93	−1.2	0.51	0.280	0.306	−0.270
Item 8	1st	55.11	−0.231	1.00	−0.1	0.45	−0.382**	0.026	0.182
	2nd	62.11	−0.558	1.00	0.0	0.43	−0.458**	0.020	0.330
Item 9	1st	57.13	−0.328	0.95	−2.0	0.51	−0.272	−0.142	−0.284
	2nd	73.60	−1.148	0.97	−0.7	0.46	−0.396	−0.182	0.298
Item 10	1st	36.48	0.637	0.96	−1.5	0.48	−0.010	−0.032	−0.220
	2nd	50.00	0.001	0.99	−0.3	0.44	−0.026	0.062	−0.082
Item 11	1st	71.83	−1.065	0.97	−0.7	0.46	−0.032	0.056	−0.132
	2nd	84.65	−1.897	1.02	0.3	0.32	−0.140	0.220	0.194
Item 12	1st	48.94	0.050	1.00	0.0	0.44	0.586**	−0.156	−0.388**
	2nd	55.92	−0.267	0.99	−0.4	0.46	0.470**	0.108	−0.294
Item 13	1st	53.23	−0.153	0.93	−2.8	0.55	−0.262	−0.258	0.158
	2nd	69.49	−0.923	0.94	−1.7	0.50	−0.052	−0.222	0.104
Item 14	1st	33.85	0.760	1.05	1.7	0.33	0.276	0.018	0.602**
	2nd	47.14	0.132	1.01	0.6	0.41	0.068	−0.034	−0.100
Item 15	1st	49.52	0.017	0.96	−1.7	0.51	−0.028	−0.030	−0.006
	2nd	72.02	−1.058	0.98	−0.5	0.45	0.054	−0.132	0.168

[a] DIF and difficulty values in logits

[b] DIF values are absolute differences between item difficulties in the respective multigroup-model, Gender: boys – girls, design: longitudinal – cross-sectional, migration: without – with **p(|DIF|=0 logits) <.05

References

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

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723. doi:10.1109/TAC.1974.1100705.

Alexander, J. M., Carr, M., & Schwanenflugel, P. J. (1995). Development of metacognition in gifted children: directions for future research. *Developmental Review, 15*, 1–37. doi:10.1006/drev.1995.1001.

Alexander, J. M., Johnson, K. E., Albano, J., Freygang, T., & Scott, B. (2006). Relations between intelligence and the development of metaconceptual knowledge. *Metacognition and Learning, 1*(1), 51–67. doi:10.1007/s11409–006–6586–8.

American Educational Research Association (AERA), American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Annevirta, T., & Vauras, M. (2001). Metacognitive knowledge in primary grades: a longitudinal study. *European Journal of Psychology of Education, 16*(2), 257–282. doi:10.1007/BF03173029.

Artelt, C., Schiefele, U., & Schneider, W. (2001). Predictors of reading literacy. *European Journal of Psychology of Education, 16*(3), 363–383. doi:10.1007/BF03173188.

Artelt, C., Beinicke, A., Schlagmüller, M., & Schneider, W. (2009). Diagnose von strategiewissen beim textverstehen [assessing knowledge about reading strategies]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 41*(2), 96–103. doi:10.1026/0049–8637.41.2.96.

Baker, L., & Brown, A. L. (1984). Metacognitive skills and reading. In P. D. Pearson (Ed.), *Handbook of Reading Research* (pp. 353–394). New York: Longman. doi:10.2466/pms.1984.59.1.159.

Belmont, J. M., & Borkowski, J. G. (1988). A group-administered test of children's metamemory. *Bulletin of the Psychonomic Society, 26*(3), 206–208. doi:10.3758/BF03337288.

Bjorklund, D. F. (2005). *Children's thinking: Cognitive development and individual differences*. Belmont: Thomson.

Borkowski, J. G., Peck, V. A., Reid, M. K., & Kurtz, B. E. (1983). Impulsivity and strategy transfer: Metamemory as mediator. *Child Development, 54*(2), 459–473. doi:10.2307/1129707.

Brown, A. L., Bransford, J. D., Ferrara, R. A., & Campione, J. C. (1983). Learning, remembering, and understanding. In J. H. Flavell & E. M. Markham (Eds.), *Handbook of child psychology: Cognitive development* (pp. 77–166). New York: Wiley.

Cattell, R. W. R., & Osterland, J. (1997). *Grundintelligenztest Skala 1 (CFT 1)* (5th ed.). Hogrefe: Göttingen.

Cavanaugh, J. C., & Borkowski, J. G. (1980). Searching for metamemory-memory connections. *A developmental study. Developmental Psychology, 16*(5), 441–453. doi:10.1037/0012–1649.16.5.441.

DeMarie, D., Miller, P. H., Ferron, J., & Cunningham, W. R. (2004). Path analysis tests of theoretical models of children's memory performance. *Journal of Cognition and Development, 5*(4), 461–492. doi:10.1207/s15327647jcd0504_4.

Ebert, S. (2014). Longitudinal relations between theory of mind and metacognition and the impact of language. *Journal of Cognition and Development*, in press.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*(4), 341–349. doi:10.1037/1040–3590.8.4.341.

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah: Erlbaum Publishers.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 8*(3), 430–457. doi:10.1207/S15328007SEM0803_5.

Ercikan, K. (2006). Developments in assessment of student learning and achievement. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (pp. 929–953). Mahwah: Erlbaum.

Flavell, J. H. (1979). Metacognition and cognitive monitoring. *American Psychologist, 34*, 906–911. doi:10.1037/0003–066X.34.10.906.

Flavell, J. H., & Wellman, H. M. (1977). Metamemory. In R. V. Kail & J. W. Hagen (Eds.), *Perspectives on the development of memory and cognition* (pp. 3–34). Hillsdale: Lawrence Erlbaum Associates.

Flavell, J. H., Miller, P. H., & Miller, S. A. (2002). *Cognitive Development* (4th ed.). Englewood Cliffs: Prentice-Hall.

Fox, A. V., & Bäumer, T. (2006). *Test zur Überprüfung des Grammatikverständnisses: TROG-D: Handbuch*. Idstein: Schulz-Kirchner.

Fritz, K., Howie, P., & Kleitman, S. (2010). "How do I remember when I got my dog?" the structure and development of children's metamemory. *Metacognition and Learning, 5*(2), 207–228. doi:10.1007/s11409–010–9058–0.

Ganzeboom, H. B., de Graaf, P. M., Treiman, D. J., & de Leeuw, J. (1992). A standard international socio-economic index of occupational status. *Social Science Research, 21*, 1–56. doi:10.1016/0049–089X(92)90017–B.

Grammer, J. K., Purtell, K. M., Coffman, J. L., & Ornstein, P. A. (2011). Relations between children's metamemory and strategic performance: time-varying covariates in early elementary school. *Journal of Experimental Child Psychology, 108*(1), 139–155. doi:10.1016/j.jecp.2010.08.001.

Hasselhorn, M. (1994). Zur Erfassung von Metagedächtnisaspekten bei Grundschulkindern [on the measurement of metamemory aspects in children]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 26*(1), 71–78.

Heller, K., & Geisler, H. J. (1983). *Kognitiver Fähigkeits-Test für 1. bis 3. Klassen (KFT 1–3)*. Beltz: Weinheim.

Joyner, M. H., & Kurtz-Costes, B. (1997). Metamemory development. In N. Cowan (Ed.), *The development of memory in childhood* (pp. 275–300). Hove, East Sussex: Psychology Press.

Justice, E. M. (1985). Categorization as a preferred memory strategy: developmental changes during elementary school. *Developmental Psychology, 21*(6), 1105–1110. doi:10.1037/0012–1649.21.6.1105.

Kline, R. B. (2005). *Principles and practice of structural equation modeling*. London: The Guilford Press.

Krebs, S., & Roebers, C. M. (2010). Primary school children's strategic regulation, monitoring, and control skills during test-taking: The role of retrieval processes and scoring scheme. *British Journal of Educational Psychology, 80*(3), 325–340. doi:10.1348/000709910X485719.

Kreutzer, M. A., Leonard, C., Flavell, J. H., & Hagen, J. W. (1975). An interview study of children's knowledge about memory. *Monographs of the Society for Research in Child Development, 40*(1), 1–60. doi:10.2307/1165955.

Kurtz, B. E., & Borkowski, J. G. (1984). Children's metacognition: exploring relations among knowledge, process, and motivational variables. *Journal of Experimental Child Psychology, 37*(2), 335–354. doi:10.1016/0022–0965(84)90008–0.

Kurtz, B. E., Reid, M. K., Borkowski, J. G., & Cavanaugh, J. C. (1982). On the reliability and validity of children's metamemory. *Bulletin of the Psychonomic Society, 19*(3), 137–140. doi:10.3758/BF03330211.

Larkin, J., & Simon, H. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science, 11*, 65–99. doi:10.1111/j.1551–6708.1987.tb00863.x.

Levin, J. R., Yussen, S. R., De Rose, T. M., & Pressley, M. (1977). Developmental changes in assessing recall and recognition memory capacity. *Developmental Psychology, 13*(2), 608–615. doi:10.1037/0012–1649.13.6.608.

Lingel, K., Neuenhaus, N., Artelt, C., & Schneider, W. (2010). Metakognitives Wissen in der Sekundarstufe: Konstruktion und Evaluation domänenspezifischer Messverfahren. *Zeitschrift für Pädagogik, 56*, 228–238.

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

Lockl, K., & Schneider, W. (2006). Precursors of metamemory in young children: the role of theory of mind and metacognitive vocabulary. *Metacognition and Learning, 1*(1), 15–31. doi:10.1007/s11409–006–6585–9.

Lockl, K., & Schneider, W. (2007). Knowledge about the mind: links between theory of mind and later metamemory. *Child Development, 78*(1), 148–167. doi:10.1111/j.1467–8624.2007.00990.x.

Martin, M. O., Mullis, I. V. S., & Chrostowski, S. J. (Eds.). (2004). *TIMSS 2003 Technical Report*. Chestnut Hill: Boston College.

McArdle, J., & Cattell, R. B. (1994). Structural equation models of factorial invariance in parallel proportional profiles and oblique confactor problems. *Multivariate Behavioral Research, 29*(1), 63–113.

Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data: an introduction. *Child Development Perspectives, 4*(1), 5–9. doi:10.1111/j.1750–8606.2009.00109.x.

Myers, M., & Paris, S. G. (1978). Children's metacognitive knowledge about reading. *Journal of Educational Psychology, 70*, 680–690. doi:10.1037/0022–0663.70.5.680.

Neuenhaus, N., Artelt, C., Lingel, K., & Schneider, W. (2011). Fifth graders metacognitive knowledge: general or domain specific? *European Journal of Psychology of Education, 26*(2), 163–178. doi:10.1007/s10212–010–0040–7.

O'Sullivan, J. T. (1993). Preschoolers' beliefs about effort, incentives, and recall. *Journal of Experimental Child Psychology, 55*(3), 396–414. doi:10.1006/jecp.1993.1022.

Paris, S. G., Cross, D. R., & Lipson, M. Y. (1984). Informed strategies for learning: a program to improve children's reading awareness and comprehension. *Journal of Educational Psychology, 76*, 1239–1252. doi:10.1037/0022–0663.76.6.1239.

Pohl, S., & Carstensen, C. (2012). *NEPS technical report - Scaling the data of the competence tests (NEPS Working Paper No. 14)*. Bamberg: University of Bamberg, National Educational Panel Study.

Pohl, S., Gräfe, L., & Rose, N. (2013). Dealing with omitted and not reached items in competence tests - Evaluating approaches accounting for missing responses in IRT models. *Educational and Psychological Measurement*. Advance online publication. doi:10.1177/0013164413504926.

Pressley, M., & Gaskins, I. W. (2006). Metacognitively competent reading comprehension is constructively responsive reading: how can such reading be developed in students? *Metacognition and Learning, 1*(1), 99–113. doi:10.1007/s11409–006–7263–7.

Pressley, M., Borkowski, J. G., & Schneider, W. (1989). Good information processing: what it is and what education can do to promote it. *International Journal of Educational Research, 13*(8), 857–867. doi:10.1016/0883–0355(89)90069–4.

Roebers, C. M., Schmid, C., & Roderer, T. (2009). Metacognitive monitoring and control processes involved in primary school children's test performance. *British Journal of Educational Psychology, 79*, 749–767. doi:10.1348/978185409X429842.

Roebers, C. M., Cimeli, P., Röthlisberger, M., & Neuenschwander, R. (2012). Executive functioning, metacognition, and self-perceived competence in elementary school children: an explorative study on their interrelations and their role for school achievement. *Metacognition and Learning, 7*, 151–173. doi:10.1007/s11409–012–9089–9.

Saß, S., Wittwer, J., Senkbeil, M., & Köller, O. (2012). Pictures in test items: effects on response time and response correctness. *Applied Cognitive Psychology, 26*, 70–81. doi:10.1002/acp.1798.

Schlagmüller, M., & Schneider, W. (2007). *WLST 7–12 - Würzburger Lesestrategie-Wissenstest für die Klassen 7–12*. Göttingen: Hogrefe.

Schlagmüller, M., Visé, M., & Schneider, W. (2001). Zur Erfassung des Gedächtniswissens bei Grundschulkindern: Konstruktionsprinzipien und empirische Bewährung der Würzburger Testbatterie zum deklarativen Metagedächtnis [Assessing metamemory in elementary school children: construction and evaluation of the Würzburg Metamemory Test]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 33*(2), 91–102. doi:10.1026//0049–8637.33.2.91.

Schneider, W. (1985). Developmental trends in the metamemory-memory behavior relationship: an integrative review. In D. L. Forrest-Pressley, G. E. Mac Kinnon, & T. G. Wallers (Eds.), *Metacognition, cognition, and human performance* (pp. 57–109). New York: Academic.

Schneider, W. (1986). The role of conceptual knowledge and metamemory in the development of organizational processes in memory. *Journal of Experimental Child Psychology, 42*(2), 218–236. doi:10.1016/0022–0965(86)90024-X.

Schneider, W. (1989). *Zur Entwicklung des Meta-Gedächtnisses bei Kindern*. Bern: Huber.

Schneider, W., & Bjorklund, D. F. (1998). Memory. In W. Damon, D. Kuhn, & R. S. Siegler (Eds.), *Cognition, perception, and language: Vol. 2. Handbook of child psychology* (5th ed., pp. 467–521). New York: Wiley.

Schneider, W., & Lockl, K. (2008). Procedural metacognition in children: evidence for developmental trends. In J. Dunlosky & B. Bjork (Eds.), *A handbook of memory and metamemory* (pp. 391–409). Mahwah: Lawrence Erlbaum Associates.

Schneider, W., & Pressley, M. (1997). *Memory development between 2 and 20*. Hillsdale: Lawrence Erlbaum Associates.

Schneider, W., Kron, V., Hünnerkopf, M., & Krajewski, K. (2004). The development of young children's memory strategies: First findings from the Würzburg longitudinal memory study. *Journal of Experimental Child Psychology, 88*(2), 193–209. doi:10.1016/j.jecp.2004.02.004.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464. doi:10.1214/aos/1176344136.

Sodian, B., Schneider, W., & Perlmutter, M. (1986). Recall, clustering, and metamemory in young children. *Journal of Experimental Child Psychology, 41*(3), 395–410. doi:10.1016/0022–0965(86)90001–9.

Sternberg, R. J. (1990). *Metaphors of the mind: Conceptions of the nature of intelligence*. Cambridge: Cambridge University Press.

Steyer, R., Eid, M., & Schenkmezger, P. (1997). Modeling true intraindividual change: true change as a latent variable. *Methods of Psychological Research Online, 2*(1), 21–33.

Swanson, H. L. (1992). The relationship between metacognition and problem solving in gifted children. *Roeper Review, 15*(1), 43–48. doi:10.1080/02783199209553457.

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring*. Mahwah: Erlbaum.

van Kraayenoord, C. E., & Schneider, W. (1999). Reading achievement, metacognition, reading self-concept and interest: a study of German students in grades 3 and 4. *European Journal of Psychology of Education, 14*, 305–324. doi:10.1007/BF03173117.

Vautier, S., & Pohl, S. (2009). Bipolarity of latent change in STAI scores. *Psychological Assessment, 21*, 187–193. doi:10.1037/a0015312.

Veenman, M. V. J., & Elshout, J. J. (1999). Changes in the relation between cognitive and metacognitive skills during the acquisition of expertise. *European Journal of Psychology of Education, 14*(4), 509–523. doi:10.1007/BF03172976.

Veenman, M. V. J., Elshout, J. J., & Busato, V. V. (1994). Metacognitive mediation in learning with computer-based simulations. *Computers in Human Behavior, 10*(1), 93–106.

Veenman, M. V. J., Van Hout-Wolters, B. H. A. M., & Afflerbach, P. (2006). Metacognition and learning: conceptual and methodological considerations. *Metacognition and Learning, 1*(1), 3–14. doi:10.1007/s11409-006-6893-0.

Von Maurice, J., Artelt, C., Blossfeld, H.-P., Faust, G., Rossbach, H.-G., & Weinert S. (2007). Bildungsprozesse, Kompetenzentwicklung und Formation von Selektionsentscheidungen im Vor- und Grundschulalter: Überblick über die Erhebungen in den Längsschnitten BiKS-3-8 und BiKS-8-12 in den ersten beiden Projektjahren. Bamberg: University of Bamberg.

Walker, C. M., & Beretvas, S. N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement, 40*(3), 255–275. doi:10.1111/j.1745–3984.2003.tb01107.x.

Weinert, F. E., & Schneider, W. (1999). *Individual development from 3 to 12: Findings from the munich longitudinal study*. Cambridge: Cambridge University Press.

Wellman, H. M. (1977). Preschoolers' understanding of memory-relevant variables. *Child Development, 48*(4), 1720–1723. doi:10.2307/1128544.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariances of psychological instruments: applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association. doi:10.1037/10222–009.

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: measuring the same construct across time. *Child Development Perspectives, 4*(1), 10–18. doi:10.1111/j.1750–8606.2009.00110.x.

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

Wu, M. L., Adams, R. J., & Wilson, M. (1997). *ConQuest: Generalised item response modelling software, Draft Release 2*. Camberwell: ACER.

Yussen, S. R., & Bird, J. E. (1979). The development of metacognitive awareness in memory, communication, and attention. *Journal of Experimental Child Psychology, 28*(2), 300–313. doi:10.1016/0022–0965(79)90091–2.

Zwick, R., Thayer, D., & Lewis, C. (1999). An empirical bayes approach to mantel-haenszel DIF analysis. *Journal of Educational Measurement, 36*(1), 1–28. doi:10.1111/j.1745–3984.1999.tb00543.x.

3.2 Manuscript 2: Linking Reading Competence Tests Across the Life Span

Original Source of Publication

Pohl, S., Haberkorn, K., & Carstensen, C. (in press). Measuring competencies across the lifespan – challenges of linking test scores. In M. Stemmler, A. von Eye, & W. Wiedermann (Eds.). *Dependent data in social sciences research: Forms, issues, and methods of analysis.* Springer.

Measuring Competencies across the Lifespan – Challenges of Linking Test Scores

Steffi Pohl,[1] *Kerstin Haberkorn*[2] *Claus Carstensen*[2]

Abstract

The National Educational Panel Study (NEPS) aims at investigating the development of competencies across the whole life span. Competencies are assessed via tests and competence scores are estimated based on models of Item Response Theory (IRT). IRT allows a comparison of test scores – and, thus, the investigation of change across time and differences between cohorts – even when the respective competence is measured with different items. As in NEPS for most of the competencies retest effects are assumed, linking is done via additional link studies in which the tests for two age groups are administered to a separate sample of participants. However, in order to be able to link the test results of two different measurement occasions, certain assumptions, such as, that the measures are invariant across samples and that the tests measure the same construct, need to hold. These are challenging assumptions regarding the linking of competencies across the whole life span. Before linking reading tests in NEPS for different age cohorts in secondary school as well as in adulthood, we, thus, investigated unidimensionality of the items for different cohorts as well as measurement invariance across samples. Our results show that the tests for different age groups do measure a unidimensional construct within the same sample. However, measurement invariance of the same test across different samples does not hold for all age groups. Thus, the same test exhibits a different measurement model in different samples. Based on our results, linking may well be justified within secondary school, while linking test scores in secondary school with those in adult age is threatened by differences in the measurement model. Possible reasons for these results are discussed and implications for the design of longitudinal studies as well as for possible analyses strategies are drawn.

Keywords: item response theory, vertical linking, competence, large-scale, measurement invariance, unidimensionality

[1] Free University Berlin, Germany.
[2] University of Bamberg, Germany.

Measuring Competencies across the Lifespan – Challenges of Linking Test Scores

Large-scale assessments generally aim at drawing inferences about individuals' knowledge, competencies, and skills (Popham, 2000). Thus, international large-scale assessments such as the Program for International Student Assessment (PISA; e.g., OECD, 2013), the Third International Mathematics and Science Study (TIMSS; e.g., Mullis, Martin, Foy, & Arora, 2012), or the Progress in International Reading Literacy Study (PIRLS; e.g., Mullis, Martin, Foy, & Drucker, 2012) aim at accurately measuring competencies, such as reading comprehension or mathematical literacy, of participants. As most of these studies have a cross-sequential design, with a new sample being drawn at every cycle, an investigation of competence development and factors influencing this development is limited. This is different in longitudinal studies, such as the National Educational Panel Study (NEPS, see Blossfeld, von Maurice, & Schneider, 2011), where due to the repeated measurement of competencies, competence development may be investigated. Specifically, the NEPS is the only study so far considering competence development across the whole life span, from newborns to adults. It does, thus, provide a rich data pool for the investigation of competence development. In order to investigate competence development, competence scores need to be linked across test administrations and test forms. While linking has so far been performed in studies across smaller age ranges, it has not been investigated whether assumptions necessary for linking test scores, hold in studies across such a long age span as in NEPS. In this study we investigated whether it is possible to link test scores for reading competence across the life span. In the following sections, we discuss the necessity of linking, describe different link designs, delineate the assumptions of linking, and discuss their plausibility in longitudinal studies. We then present the National Educational Panel Study and derive specific research questions.

Linking of Test Scores

Necessity of Linking Test Scores

It is often not feasible to administer the same test to the participants across time or age, but tests need to be adapted in difficulty and content to the respective age group. Thus, a direct comparison of competence scores from different tests is not possible, since differences in competence values across different tests represent both, differences in competence *and* difference in test items. In longitudinal studies, it is a major aim to investigate competence development over time or to compare the competencies of different age cohorts. In order to be able to compare competence scores across time or cohorts from different tests assessing the same dimension, the test scores need to be linked.

As described by von Davier, Carstensen, and von Davier (2008) and Kolen and Brennan (2004) linking means to establish a common scale for different measurement instruments that are intended to measure the same construct. Vertical linking allows for placing the competence scores of different test forms for different age groups on the same scale, thus allowing for a comparison of these test scores. IRT provides means to develop vertical scales encompassing different test versions. In order to obtain a common scale, certain test designs and analyses methods are necessary.

Link Designs

For linking of test scores, some common information or overlap between different test administrations (say grades) to be linked is needed. This can be achieved by various linking designs (for an overview see, e.g., Kolen & Brennan, 2004, Reckase, 2009, or von Davier et al., 2008). Overlap can be achieved by collecting common observations in a) a common-person design b) a common-item design, or c) a scaling-test design. In a common-person design a sample of subjects takes the two test forms to be linked. Because of the single group completing

both tests, differences in the scores on these tests can be attributed to differences in the test forms. In a common-item design two samples of different populations take two tests and the link is established by a set of common items within both tests (anchor items). This design is also called the nonequivalent group anchor test (NEAT) design (e.g., Reckase, 2009; von Davier et al., 2008). Assuming invariance of item functioning (i.e., no item drift), the common items may be used as anchors for establishing a common scale between the test versions. In vertical linking, the common-item design with overlapping items is often used across adjacent grades. The scaling-test design can be seen as a special form of the common-item design. Whereas in the common-item design, anchor items are usually administered across adjacent grades, in a scaling-test design, a common test, appropriate to all levels of ability, is implemented in each grade in addition to grade-specific items. Consequently, all students of a study deal with the same test and additionally answer items specifically constructed for their age group. There are different challenges associated with each of these designs which have to be considered (Kolen & Brennan, 2004). For instance, in the common-item or scaling-test design, one has to assume that there are no retest effects. Otherwise, item drift might occur and the measurement model would change. There is no such threat in the common-person design; instead this design requires drawing an additional sample, which is less economic, and the challenge arises from an adequate sampling strategy. Note that it is also possible to combine the different designs to build more complex data collection designs (see also Dorans, Pommerich, & Holland, 2007; von Davier, Holland, & Thayer, 2004).

Coherence of Measurement

Assumptions for Linking

In order to establish a link between test forms that allows one to depict change across time or cohorts, certain assumptions need to be fulfilled (e.g., Camilli, Yamamoto, & Wang, 1993;

Doran & Cohen, 2005; Hoover, 1984; Linn, 1993; Mislevy, 1992; Tong & Kolen, 2007): the construct to be measured needs to be the same across a) samples and b) tests. This implies that a) measurement invariance of the same items in different samples holds and that b) the items of two different tests form a unidimensional construct. Violations of these assumptions may lead to errors in linking (Monseur & Berezner, 2007; Monseur, Sibberns, & Hastedt, 2008). As a consequence, change scores do not only represent competence development but also changes in the test instrument and inferences on competence development or cohort differences will be biased.

Plausibility of Assumptions in Empirical Studies

Some researchers have stated that the assumption of measuring the same construct is hardly met in applications (e.g., Martineau, 2006; Reckase & Martineau, 2004; Wang & Jiao, 2009). For instance, Wu (2010) reported that 'In general, the further the grades are apart the less reliable the vertical scaling across grades is found to be' (p.23). We draw on studies assessing competencies that incorporated longitudinal or multi-cohort designs for collecting evidence on whether and how coherent measurement of competencies may be obtained. We first reviewed studies that Kristen, Römmer, Müller, and Kalter (2005) found in a systematic stocktaking of the most important longitudinal studies on educational pathways in selected countries in Europe and North America. Kristen et al. identified a number of longitudinal large-scale studies in education. These usually considered competence assessment across some part of the life span. Only a few of them included competence assessment in their design and for those who did hardly any information on vertical scaling and on tests of assumptions of linking was available. For those that did assess competencies, results on the coherence of measurement were ambivalent. Additionally to the studies reviewed in Kristen et al., we collected information on measurement coherence from small-scale studies or multi-cohort studies.

Evidence supporting coherence of measurement

There are some studies that did find evidence for the coherence of measurement. One of them is the National Education Longitudinal Study of 1988 (NELS: 88; Rock, Pollack, Owings, & Hafner, 1991), a very prominent longitudinal study on competencies in the USA. In this study students were followed in intervals of two years from 8th grade to 24-25 years. For the three waves of data collection in school in 8th, 10th, and 12th grade, students' reading, math, social studies, and science competencies were assessed (Rock, Pollack, & Quinn, 1995). In order to link the test forms of the competence tests across age, a common-item design was used. Half of the items (in reading) to three quarters of the items (in math) from one measurement occasion were also used in the following assessment. The authors reported that measurement invariance was found across measurement occasions.

Another longitudinal study for which a coherent measurement was supported is the Early Childhood Longitudinal Study (ECLS; Pollack, Atkins-Burnett, Najarian, & Rock, 2005) in the USA. It consists of a birth cohort (ECLS-B), with measurements starting with 9 month old children which are followed up to 1st grade, and two kindergarten cohorts (ECLS-K and ECLS-K:2011), one ranging from fourth to eighth grade and the second following children from kindergarten till fifth grade. In the kindergarten cohorts reading, math, and scientific competencies were assessed and linking was performed using a common-item design. Analyzing differential functioning of the items in the ECLS-K study across time, Pollack and colleagues (2005) found measurement invariance of the common items across measurement occasions. Thus, in this study measurement invariance across the wide span from kindergarten to secondary school could be assured.

Besides these large-scale studies, there is some evidence on the coherence of competence measures across age from other studies. Wang and Jiao (2009), for example, investigated the

equivalence of the factorial structure of the Stanford Reading Comprehension Test (Stanford Achievement Test Series, Tenth Edition, 2004) across eight samples in grades 3 to 10. They found that on subtest-level the measurement models were invariant across grades. While Wang and Jiao investigated measurement invariance only on subtest level, in a longitudinal study, Wang, Jiao, and Zhang (2013) investigated measurement invariance of the Measures of Academic Progress (MAP) for mathematic and reading competence on item level. The authors found that measurement invariance could be assured across 5th to 7th Grade.

Evidence questioning coherence of measurement

However, there is also evidence that the competence assessed changes across time or cohorts. This is the case in the BiKS-3-10 study on Educational Processes, Competence Development, and Selection Decisions at Preschool and Elementary School Age (von Maurice et al., 2007), a longitudinal study on competence development and educational progress from kindergarten to primary school. Linking between testing waves was done via a common-item design. Robitzsch, Dörfler, Pfost, and Artelt (2011) investigated measurement invariance of the common items of a reading competence test between three measurement occasions between Grade 3 and Grade 4. The authors found considerable item drift across measurement occasions, threatening the interpretation of change scores as indicators of competence development.

Also some cross-sectional large scale studies, specifically the National Assessment of Educational Progress (NAEP) and a German study evaluating the National Educational Standards (NES) found evidence for measurement non-invariance across age. The National Assessment of Educational Progress (NAEP), the largest representative educational assessment in the USA, explores achievement of students in various domains, among others mathematics and reading, every two years in Grades 4, 8, and 12 (Jones & Olkin, 2004). After the first waves of assessments, measurement invariance of anchor items across grades was checked and threats to

measurement invariance were reported on a significant number of mathematics and history items (Haertel, 1991; MCClellan, Donoghue, Gladkova, & Xu, 2005), whereas the reading test functioned rather well across grade levels. Altogether, Haertel questioned the usefulness of cross-age scales for the NAEP regarding the costs in terms of constraints on the framework. He even concluded that comparing students separated by four to eight years is 'largely meaningless' (p. 14). As a consequence in the following assessments cross-age comparisons were discouraged (Thissen, 2012). Threats to measurement invariance were also found in the evaluation of the German National Educational Standards (NES; Klieme et al., 2003; Rupp & Vock, 2007) by the Institute for Educational Progress (IQB). In the domain of language assessment, Böhme and Robitzsch (2009) analyzed reading tests of pilot and calibration studies which were administered in a cross-sectional setting in Grade 3 and 4 of elementary school. For evaluating the item parameter drift, the authors evaluated the variance of differential item functioning (DIF) between the two grades. DIF occurs when items function differently for different groups, that is, when estimated item difficulties differ between subgroups after controlling for overall group differences on the latent trait. Based on the classification scheme of Penfield and Algina (2006), the results indicated a medium DIF variance and some items considerably favored third or fourth graders.

In addition to the above mentioned large-scale studies, we also reviewed small longitudinal studies. As such in a study on science competence development, Carstensen, Lankes, and Steffensky (2012) found that measurement invariance was not warranted for common science items across three measurement occasions in 5^{th} to 6^{th} year old children. In an U. S.American study, Tong and Kolen (2007) investigated the performance of various vertical linking methods in simulation studies as well as empirical data. The analyses of the empirical data were based on the assessments of the Iowa Tests of Basic Skills (ITBS; Hoover, Dunbar, & Frisbie, 2003) in the

four different domains vocabulary, mathematics, language, and reading covering Grade 3 through Grade 8 via a scaling- and an anchor-test design. Tong and Kolen found that the scaling designs in the empirical studies produced scales with dissimilar properties, especially for tests that tended to be less homogeneous in content across grades and for tests that included testlet-based items such as the reading test.

Summary of previous findings on coherence of measurement

The results from previous longitudinal or multi-cohort studies show that the assumption of measuring the same latent variable across different age groups is not a trivial one. Indeed, results of some studies such as NELS or ECLS-B confirmed measurement invariance across age, but other studies such as NAEP or BiKS report challenges in creating a common scale. Even in studies with a short age span such as in the NES study or the study by Carstensen et al. (2012) measurement invariance is not always fully warranted. The issue of coherence of measurement is even more prevalent in the NEPS covering such a broad age span.

The National Educational Panel Study - Competence development across the life span

The German National Educational Panel Study (NEPS, see Blossfeld et al., 2011) is a current longitudinal study on competence development in Germany. A particular strength of the NEPS is that it considers competence development and educational pathways across the whole life span. NEPS incorporates a multicohort sequence design (see Figure 1) that incorporates around 60.000 target persons in six different starting cohorts (newborns, children in kindergarten, students in fifth grade, students in ninth grade, university students, and adults). In order to provide information on educational processes already at an early stage of the study, the six

starting cohorts simultaneously started in 2010[1] at different important educational stages and are followed concurrently in their development over time.

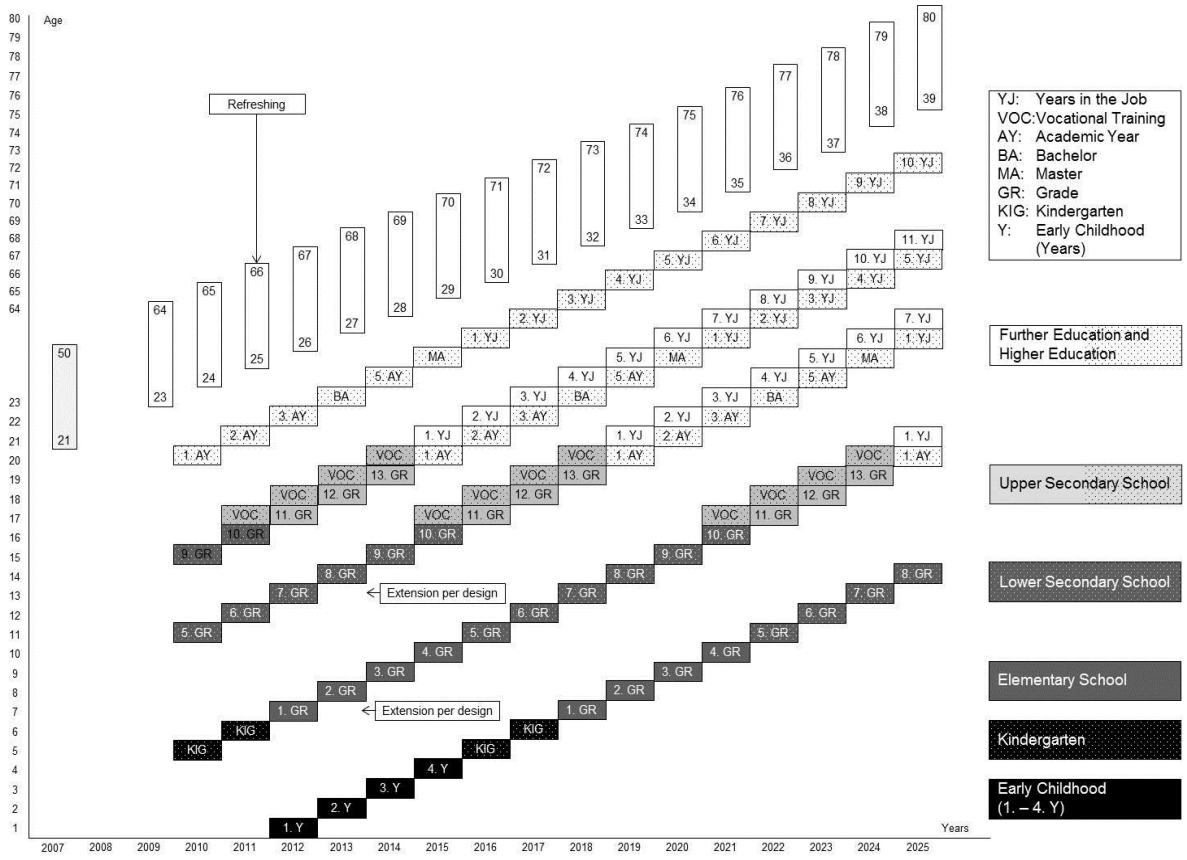


Figure 1. The multi-cohort sequence design of the NEPS.

By regarding different cohorts that overlap at some point in the design, it is possible to investigate educational processes across the whole life span without following the same participants across their whole life. [2] Competencies as well as a variety of data on conditions for and consequences of individual educational careers are assessed. Information is gained from the target persons as well as their parents, teachers, or other educators. For many cohorts, different competence domains are repeatedly measured every two years, allowing researchers to explore

[1] Newborns started 2012 and the adult sample was pursued from the former ALWA study.
[2] This is possible if measurement invariance for the instruments for comparisons between cohorts can be assumed. One may also investigate and account for cohort effects with this design.

the evolvement of these competencies. Based on the data 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 et al., 2011).

Competence Assessment in the NEPS

The framework for assessing competencies in the NEPS employs a number of different domains (Artelt, Weinert, & Carstensen, 2013). These include, among others, reading competence (Gehrer, Zimmermann, Artelt, & Weinert, 2013), mathematical competence (Neumann et al., 2013), scientific literacy (Hahn et al., 2013) and information and communication technologies (ICT) literacy (Senkbeil, Ihme, & Wittwer, 2013). The NEPS aims at assessing these domain-specific competencies coherently across the life span in order to appropriately describe the competencies' developmental progress over time. Therefore, competence models have been specified comprising a consistent structure of each domain across ages and cohorts (Weinert et al., 2011). In order to facilitate a coherent competence assessment, the same conceptual framework has been applied for the tests of different age groups.

For reading competence, for instance, the same cognitive processes and text types are used in the tests across different age groups. According to the competence models, new tests are developed and evaluated in pilot studies with the most appropriate items being used in the final main study tests in the NEPS. The newly developed test instruments require the participants to respond to tasks with different response formats. The responses to these tasks are scaled using models of Item Response Theory. In the NEPS, reading, mathematical, scientific, and ICT competence are scaled using the Rasch (Rasch, 1960) or the Partial Credit Model (Masters, 1982) (for the scaling model in the NEPS see Pohl & Carstensen, 2012). Here, we focus on the measurement of reading competence from Grade 5 to adulthood.

Linking in the NEPS

Linking in the NEPS includes linking of test scores within cohorts over measurement occasions as well as across cohorts. Linking test scores within cohorts is obviously needed to enable the analysis of change over time within each cohort of the NEPS. For example, a question might be how reading competence of students develops between fifth grade (in 2010) and ninth grade (in 2014). However, with the multi-cohort-sequence design of the NEPS, comparisons between cohorts are also intended. As an example, a question might focus on how much ninth graders in 2010 differ in their reading competence from fifth graders, at the same measurement occasion.

In the NEPS different linking strategies are employed. Since retest effects are expected for reading and science items, neither a common-item nor a scaling-test design are applicable as the same items would need to be presented twice to the participants. Instead, common person designs were employed to obtain linking information. In the NEPS, link samples are additionally drawn randomly from the older of the two age groups, that is, students in the link sample typically take the on-grade and below-grade test. In the domain of mathematical competence retest effects are not expected and both common-item designs as well as common-person designs are implemented (Pohl & Carstensen, 2013).

Coherence of Measurement in NEPS

Coherence of measurement is a special challenge in the NEPS as the NEPS, in contrast to many other educational studies, follows the development of persons across the whole life span. In constructing the test instruments a great deal of effort was put on a coherent assessment of competencies over the life span (see, e.g., Gehrer et al., 2013; Neumann et al., 2013, or Hahn et al., 2013). For (almost) all age groups the same conceptual framework, the same cognitive demands, as well as the same item formats were used for test construction. However, while the

assumption of comparable test scores seems to be very plausible for cohorts that are similar in age and in educational or institutional setting, such as linking Grade 5 to Grade 7 students, it is still questionable across 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 settings (Grade 9 students being in school and used to tests and adults mainly being in labor market). This poses a challenge on the comparability of test results across time and cohorts.

Possible threats to the assumptions of linking (Camilli et al., 1993; Hoover, 1984; Tong & Kolen, 2007) were addressed in the NEPS. In the NEPS fixed item position within a test and rotation of test position within a testlet were used to control for position effects. The possible mismatch of item difficulty to person ability was evaluated in pilot studies within test development and was assured for the main samples. Note that they do, however, not necessarily need to hold for link samples. From the construction point of view, in the NEPS effort is invested to assure coherent measurement of the same latent variable across age groups. Whether this proves successful needs to be tested empirically.

Research Questions

The NEPS is the first study that aims to measure competencies across *the whole life span.* So far, there has been no empirical evidence whether and how coherent measurement may be obtained across such a wide age range. The aim of the present study was to investigate whether the construction of coherent instruments for the measurement of competencies across the life span is possible and as such was successful in the NEPS. Here we focused on reading competence and investigated whether it is possible to measure reading competence *coherently* from fifth grade to adulthood. Specifically we asked whether the assumption holds that the

measured reading competence is the same for different age cohorts and measurement occasions. Methodologically phrased, we investigated whether the assumptions for vertical scaling are met, that is: 1) Is competence measurement on reading invariant across studies and age groups? and 2) Is reading competence in NEPS unidimensional across age groups? Additionally we explored item and test characteristics related to the coherence of measurement. Only if a competence measurement is coherent and an adequate link between measurements can be established, we may investigate development and change of the competencies (which is one of the main aims in longitudinal studies) as well as compare competencies across different cohorts (on which the multicohort sequence design relies).

Method

Sample and Design

Sample

In the present study we analyzed data from four main studies (in Grade 5, Grade 7, Grade 9, and on Adults) and three corresponding link studies of the NEPS. The three link studies are designed to link the measurements of the main studies between Grade 5 and Grade 7 (G5-G7), between Grade 7 and Grade 9 (G7-G9), and between Grade 9 and Adults (G9-AD). Thus, the studies considered in this paper allow for linking reading competence measures from Grade 5 to adults. The main study in Grade 5, Grade 9, and on Adults took place in the first assessment wave of the NEPS (starting in 2010). The subjects in these studies comprised different starting cohorts. The second competence assessment of the fifth graders of 2010 took place in 2012 in Grade 7. As the main studies of Grade 5 and Grade 7 comprised the same starting cohort, most of the subjects in Grade 5 also participated in the assessment in Grade 7. The link studies were administered parallel to the last of the main studies that are to be linked. Thus, the link study G9-AD took

place in the first wave of the NEPS in 2010, while the link studies G5-G7 and G7-G9 were carried out in 2012 (when the main study in G7 took place). The participants in the link studies were always drawn from the older of the two populations, e.g., for linking Grade 9 students to adults, the link study was performed on an adult sample.

The main studies had sample sizes between 5000 (in Grade 5 and Adults) up to about 14000 (in Grade 9) participants, whereas the link samples were considerably smaller with 500 to 600 participants (see Table 1). In all main studies, the participants constituted representative samples of German inhabitants at different ages (Aßmann et al., 2011). For the link study G9-AD, adults were representatively drawn from the 16 German federal states, while the link studies G5-G7 and G7-G9 were conducted in only four federal states: Lower Saxony, Bremen, North Rhine-Westphalia, and Saxony. Although no representative sample of the whole country could be drawn for two of the link studies, representative samples were drawn from the four federal states and we did not expect large differences in populations. However, it is to note that participants in the main studies agreed to take part in a longitudinal study, while participants in the link study were only recruited for one assessment. This may result in different participation processes and, thus, in different populations.

Looking at demographic characteristics (Table 1), the link studies and the respective main studies seem to be rather similar. Comparing the main study in G7 with the link study G5-G7, a relatively equal distribution of male and female students and similar percentages of school type and migration background were found when missing values were not taken into account. The average age in the link study G5-G7 was almost identical to that in the main study G7. Based on the design, students in the main study G5 were about two years younger than in the corresponding link study G5-G7. They were, however, similar in many of the other demographic characteristics.

The link study G7-G9 and the corresponding main studies in G7 and G9 featured similar properties regarding gender and migration background. However, the link study G7-G9 and the main study G9 slightly differed in age, with the participants in the main study being on average about half a year older. Participants in the different studies also differed in school type. There were more students in the highest academic track in the main study in G7 than in the link study; the lowest number of students in the highest academic track was found in the main study in G9. Thus, the link study G7-G9 and the respective main study in G9 may have been drawn from different populations.

Adults in the main study and the corresponding link study G9-AD had a similar age distribution and a similar percentage of persons with migration background. Slight differences occurred on the variables gender and school degree. These differences possibly reflect differences in participation between the two studies. The Grade 9 students in the main study and adults in the link sample G9-AD were by design drawn from different populations and they differed in some of the background variables. Note that in the school cohorts the dichotomous variable school type/degree refers to the school type participants attend at the moment. For the school cohorts the variable differentiates between students attending grammar school (German: Gymnasium) and students with a lower school type. Since (most of the) participants in the Adults sample did not attend school any more, the respective variable refers to the highest school degree achieved so far, distinguishing between an A-level degree (German: Abitur) and a lower school degree. In the current study, the Grade 9 sample and the link study sample differed in the distribution of school type/degree, and additionally in the variables gender, and migration background. Moreover, as expected by design, the link study sample was substantially older than students in Grade 9.

In summary, while the link study G5-G7 shows similar demographic properties as the corresponding main studies G5 and G7, there are some differences in the samples between the

link study G7-G9 and the corresponding main studies as well as the link study G9-AD and its corresponding main studies.

Table 1. Description of the samples in main and link studies

	Main study **G5**	Link study **G5-G7**	Main study **G7**	Link study **G7-G9**	Main study **G9**	Link study **G9-AD**	Main Study **Adults**
N	5193	608	6186	534	13897	502	5335
Gender (rel. freq.)							
Male	51.6%	48.6%	51.7%	51.1%	50.2%	43.5%	49.9%
Female	48.4%	51.4%	48.3%	48.9%	49.8%	56.5%	50.1%
Age Mean (SD)	10.9 (0.5)	12.9 (0.6)	13.0 (0.5)	15.3 (0.7)	15.7 (0.6)	45.2 (12.7)	47.6 (10.9)
Migration background (rel. freq.)							
No	68.0%	69.7%	66.6%	71.5%	70.5%	83.7%	80.3%
Yes	25.1%	26.0%	22.0%	23.2%	25.0%	15.5%	14.6%
No information	6.9%	4.3%	11.3%	5.2%	4.5%	0.8%	5.2%
School type/degree (rel. freq.)							
Lower school type	54.3%	56.0%	53.0%	59.1%	65.0%	66.0%	54.6%
High school type	45.4%	44.0%	47.0%	40.9%	35.0%	34.0%	45.4%

Migration background either the person itself or one of its parents is born in a foreign country; *School type/degree* refers to the school type in the school cohort samples and to the school degree in the Adults samples; high school type: at least grammar school/A-level degree, lower school type: other school types/a lower school degree.

Design

A common-person link design was used to link the reading competence scores of different age groups. We describe the design exemplary for linking the Grade 9 test to the adult reading test. The link sample was always drawn from the older population of the two main studies to be linked. Thus, the link study G9-AD was conducted on adults. In the main studies, one test constructed for this age group was administered, while the link sample completed the tests of the two adjacent years. Regarding the link between Grade 9 and Adults, the 9th graders and the adults

in the main studies received only the Grade 9 test or the Adults test, respectively. In the link study, both tests were administered to the participants. The two tests in the link studies were given in randomized order to balance position effects. The same link design was applied for linking competence scores of Grade 5 students to Grade 7 students and of Grade 7 students to Grade 9 students.

Whereas in the first testing wave (here main studies in Grade 5, Grade 9, and Adults), reading competence was measured using a single test form for all students, in later waves (here Grade 7) longitudinal multi-stage testing using information from the previous testing wave for routing to test forms of different difficulty was applied in order to enhance test targeting, motivation, and measurement precision (see Pohl, 2014). Thus, the test in Grade 7 consisted of two test forms that differ in mean difficulty. 61.9 percent of the students in Grade 7 took part in the previous competence testing wave in Grade 5, so competence scores from the previous wave were available for these students. Additionally, 2357 (38.1%) new students were recruited in Grade 7 to enlarge the sample size. Students with an ability estimate in Grade 5 below the median were assigned to an easy test form in Grade 7 (N=1771), students with an ability estimate equal or greater than the median were assigned to the difficult test form (N=2058) (see Figure 2). Students with no available competence score from Grade 5 (N = 2357) were assigned to the difficult test form, since pilot studies had shown that the difficult test form targets a wider ability range than the easy test form. Altogether, 1771 students in the main study in Grade 7 took the easy test form, and 4415 subjects took the difficult test form. The assignment to the different test forms was different in the corresponding link studies (G5-G7 and G7-G9). As these were cross-sectional samples, no preliminary information about the student's competencies was available and the two test forms of the G7 reading competence test were administered randomly to the participants of the link studies. Note, that the different assignment of test forms results in

different population characteristics between the main study and the link studies, conditional on the test

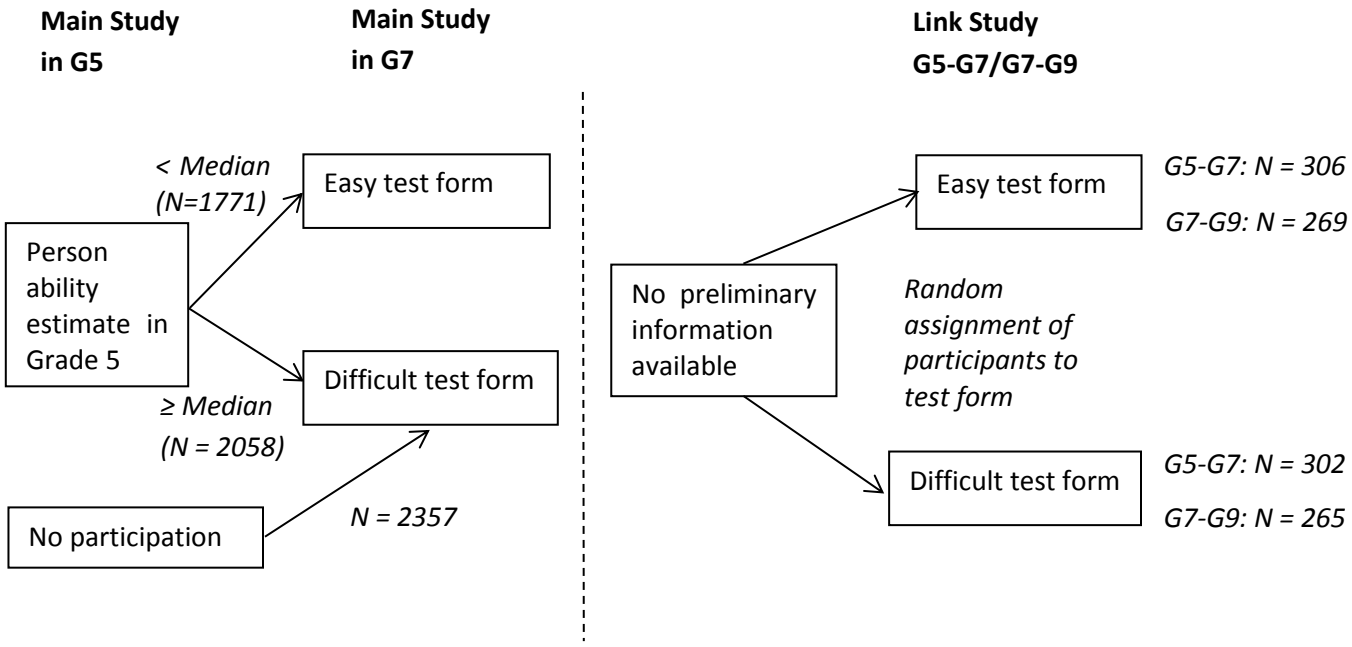


Figure 2. Allocation of the Grade 7 test forms to the examinees in the main study and the link studies.

Measures and Procedures

In the NEPS, reading competence tests were developed that aim at measuring reading competence coherently across the life span (Weinert et al., 2011) – using the same conceptual framework across age (Gehrer et al., 2013). The NEPS framework on reading competence embodies different text functions and different cognitive requirements. Tests across all ages consist of five texts each with a different text function: 1. information texts, 2. commenting or argumenting texts, 3. literary texts, 4. instruction texts, and 5. advertising texts. The specific questions focusing on the texts' content can be classified into three types of items according to their cognitive requirement: (a) finding information in texts b) drawing text-related conclusions, and c) reflecting and assessing. The three types of items are not intended to primarily differ in

difficulty, but qualitatively (Gehrer et al., 2013). Most of the items are multiple-choice (MC) items with one option out of four being correct. Furthermore, complex multiple choice (CMC) items and matching (MA) items are included in the tests. CMC items consist of several subtasks with two response options, MA items include a number of responses which have to be matched to a given set of statements. Subtasks of CMC and MA items were aggregated to one polytomous variable per item and given (partial) credit scores (see Haberkorn, Pohl, Carstensen, & Wiegand, submitted; Pohl & Carstensen, 2012).

As described above, in Grade 7 two test forms were administered which differed in difficulty, and students were assigned to either the difficult or the easy test form. Each test form comprised five texts, and the two test forms had three out of five texts (plus the respective items) in common which enabled a link between the test forms. The three common texts were presented on the same positions in both test forms.

In the main and link studies the reading competence test was administered with other competence tests and questionnaire items assessing further information of the examinees. The reading test featured a paper-and-pencil format and participants had 30 minutes to complete the test. While the test was presented to the students in Grade 5, 7, and 9 in a group setting at school with a group size of up to 25 subjects, the adults took the test individually at their homes.

Analyses

We scaled the data within the framework of Item Response Theory (IRT). As described above, the reading test included simple MC, complex MC and matching items. The complex MC and the matching items consisted of a set of subtasks that were aggregated to a polytomous variable in the final scaling model in the NEPS. In accordance with the scaling procedure for competence data in the NEPS (Pohl & Carstensen, 2012, 2013), we used the Partial Credit model (Masters, 1982) for scaling the data. The models were fitted to the data using ConQuest (Wu,

Adams, Wilson, & Haldane, 2007). Missing responses were ignored in the estimation of the parameters (see Pohl, Gräfe, & Rose, 2014).

We evaluated both assumptions of measurement invariance. First, we investigated the dimensionality of the tests. For this, we used the link samples, which took the reading tests of two adjacent age groups, and specified a) a two-dimensional model – each test form of a specific age group forming one dimension and b) a unidimensional model across both tests. For the Grade 7 test, both test forms were included in the analyses and the information of test form assignment was included in the model. The dimensionality of the test was assessed by comparison of the AIC and BIC of the two models and by evaluating the latent correlation between the dimensions of the two test forms (estimated in the two-dimensional model). If the model comparison supports a unidimensional model and the correlation between the test forms is close to one, the assumption that the tests of adjacent age groups measuring the same construct within one population is supported.

Second, we investigated whether the tests measure the same construct in the different studies. For this purpose, we applied a multi-group Rasch model and evaluated differential item functioning (DIF) by comparing estimated item difficulties between main study and link study. Note that for the tests, where main study and link study are drawn from the same population, the test of DIF is mainly a test for equivalence of the samples drawn. DIF of items that were administered to different populations in the main study and the link study is mainly a test of measurement coherence across age groups and settings. For the Grade 7 test, DIF was investigated separately for the easy and the difficult test form. Although the participants in the main study G7 and the link study attend the same grade, differences in populations are present, as the assignment to the different test forms differed between main study and link study. The

populations may especially differ in person abilities and as a consequence possibly also in test taking strategies.

In subsequent analyses we investigated whether there is a relationship between DIF and test as well as item characteristics as possible explanations for measurement variance. We specifically considered the competence domain assessed, item difficulty, text functions, cognitive requirements, and response format.

Results

Dimensionality

Using the link studies we investigated whether the reading tests of adjacent years do measure a unidimensional construct. The results showed that in all three studies the fit indices supported a two-dimensional over a unidimensional model (see Table 2). It is, however, to note that the differences in AIC and BIC were rather small compared to sample size and test length, so that statistical inferences will not be without ambiguity (Alexandrowicz, 2008). The latent correlations between the test forms of two adjacent age groups were very high (see Table 2), indicating that within the same sample, the different tests measure the same construct.

Table 2. Fit indices of the uni- and the multidimensional models in the link studies

Link Study	Model	AIC	BIC	Latent correlation
G5-G7	unidimensional	32782.13	33267.25	
	two-dimensional	32756.85	33250.79	0.93
G7-G9	unidimensional	27013.45	27462.89	
	two-dimensional	26993.14	27451.14	0.95
G9-AD	unidimensional	25257.80	25586.85	
	two-dimensional	25241.78	25579.26	0.95

Measurement Invariance

In the following the results on measurement invariance are presented by reporting the DIF between main study and link study for each of the three links. Afterwards the relationships of item and test characteristics with DIF are described.

Linking Grade 5 to Grade 7

The absolute differences in the estimated item parameters of the Grade 5 test in the main study of Grade 5 and the link study G5-G7 are presented at the top of Fig. 3. Note that the test was administered to Grade 5 students in the main study and to Grade 7 students in the link study and, thus, allows one to describe differences in item functioning across age groups. As can be seen in the Figure, the differences in item difficulties between the two studies were negligible, ranging from -0.794 to 0.504 logits. For only one item DIF exceeded 0.6 logits. Overall, the measurement model of the Grade 5 test in the main study on Grade 5 students seems to be similar to that in the link study on Grade 7 students. In Fig. 3 also DIF for the items of the easy and the difficult Grade 7 test is shown. Although the main study and link study were both sampled from the population of Grade 7 students, the assignment to test forms differed between main study and link study. In the main study the assignment was based on ability estimates from the previous testing waves, resulting in subgroups with a rather homogenous ability and a good test targeting. In contrast, random assignment was performed in the link study, resulting in heterogeneous subgroups and a test targeting that was less tailored to the ability level of the subgroups. As in the main study of Grade 7 the students newly recruited in Grade 7 all received the difficult test (regardless of their ability), the competence distribution for the students receiving the difficult test should be more similar between main study and link study than for the easy test form. This is also reflected in the results of measurement invariance of the test forms across samples (Fig. 3). DIF was smaller for the difficult test form than for the easy test form. DIF values ranged from -

0.606 to 0.480 logits in the difficult test form and from -0.664 to 0.750 logits in the easy test form. Only one item in the difficult test form and four items in the easy test form showed DIF greater than 0.6.

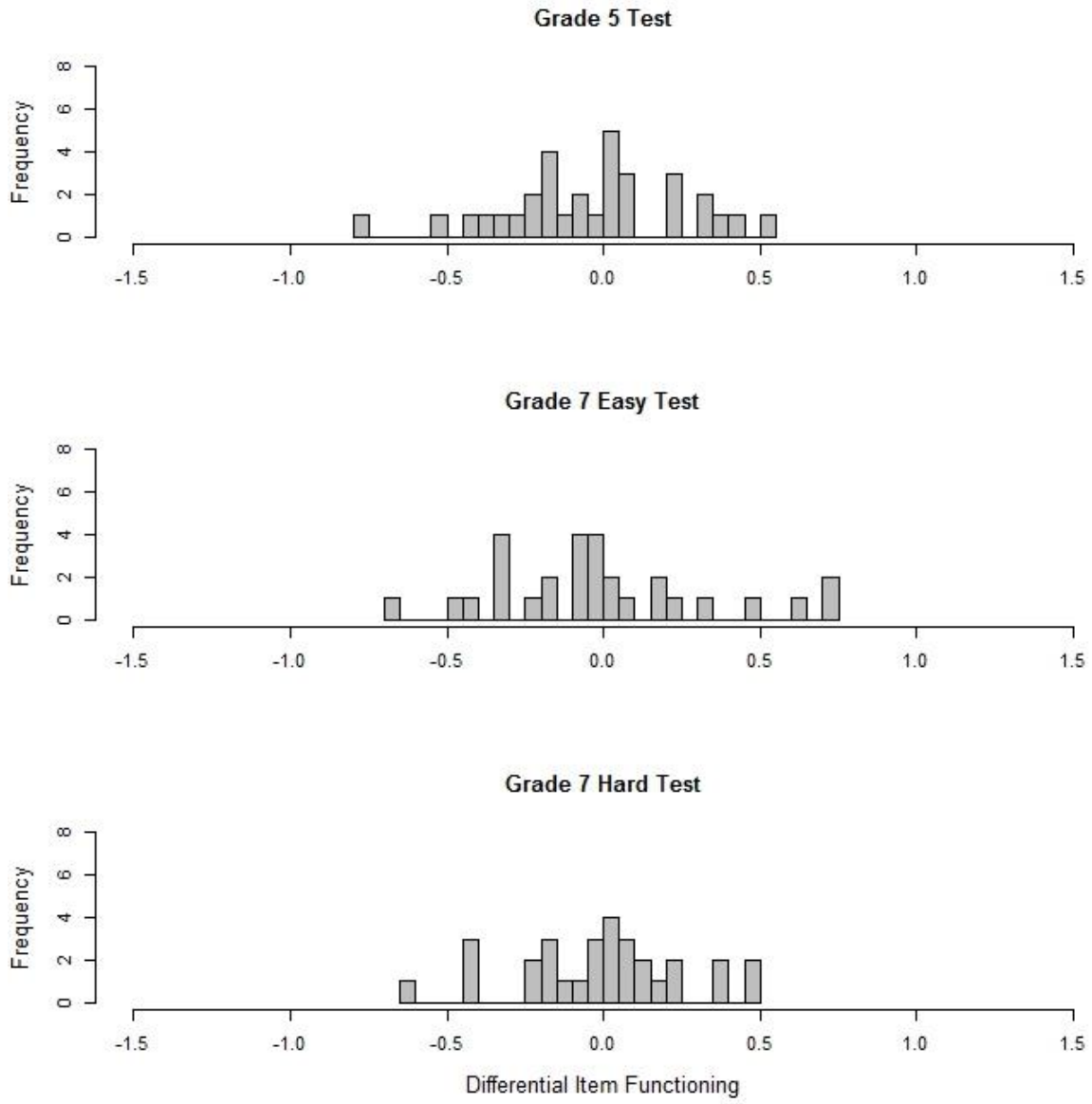


Figure 3. DIF of items linking Grade 5 to Grade 7.

Linking Grade 7 to Grade 9

The results on measurement invariance linking the tests in Grade 7 to the test in Grade 9 are presented in Fig. 4. There was no considerable DIF for the items of the Grade 9 test. For all

items absolute differences in estimated item difficulty were less than 0.5. As for the Grade 9 test, the samples of the main study and the link study were both drawn from the population of 9th graders, these results support the comparability of the samples.

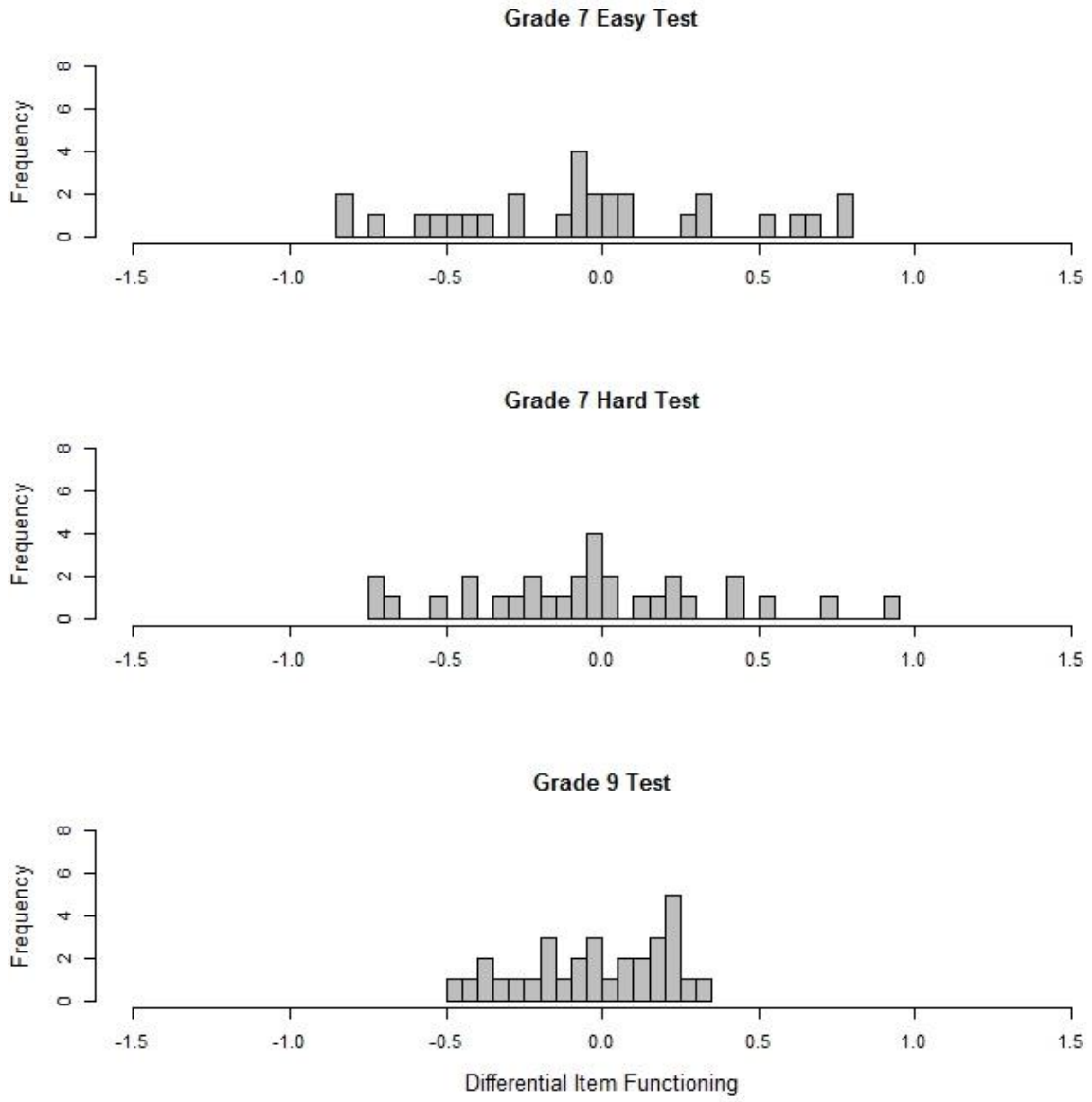


Figure 4. DIF of items linking Grade 7 to Grade 9.

Considering the link across different age groups, there was noticeable DIF for both test forms in Grade 7 across samples. DIF values ranged from -0.846 to 0.792 logits in the easy test form and from -0.746 to 0.914 in the difficult test form. There was also a non-negligible amount

of items with rather large DIF, especially in the easy test form. For seven items in the easy and five items in the difficult test form DIF exceeded 0.6. The results indicate that the two test forms function differently in the different populations.

Linking Grade 9 to Adults

Figure 5 shows the differences in estimated item difficulties for linking the Grade 9 test to the adult test. For the adult test, estimated item difficulties were very similar across the main study and the link study, indicating similarity of both samples. No DIF value exceeded an absolute value of 0.4 logits (range from -0.300 to 0.392 logits). This was different for the Grade 9 test, where main sample and link sample were drawn from different populations.

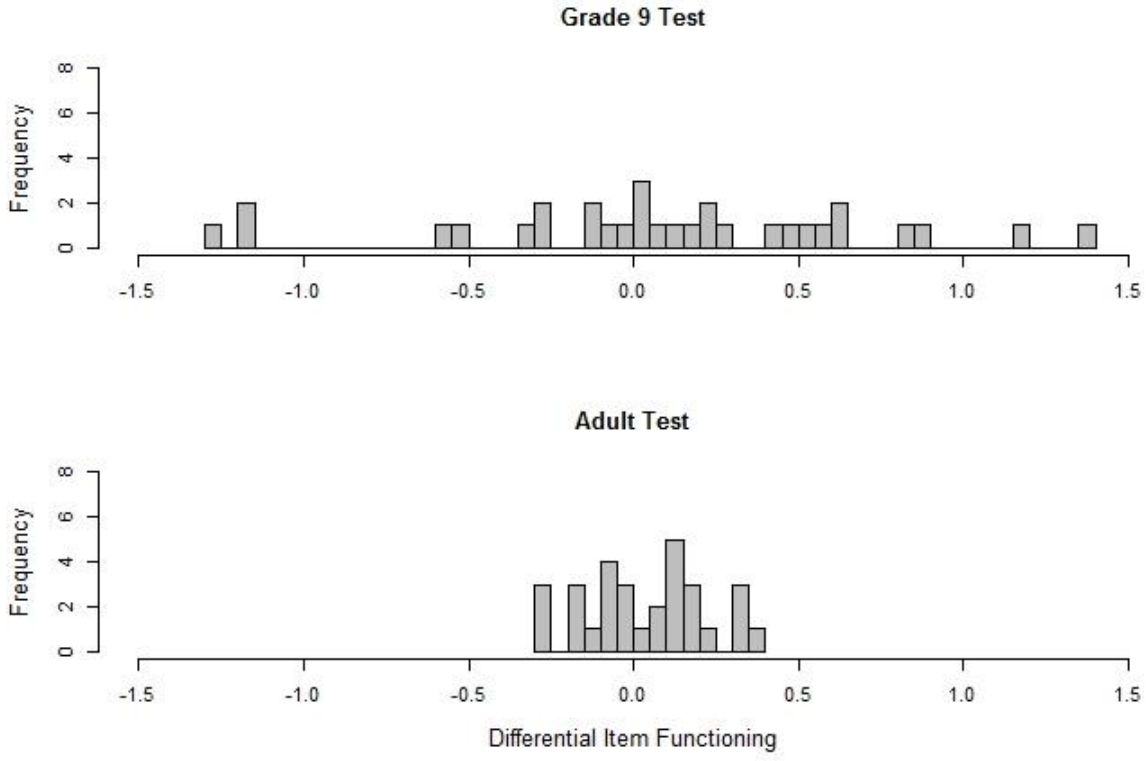


Figure 5. DIF of items linking Grade 9 to Adults.

DIF values were large, ranging from -1.298 to 1.394 logits. Nine items exhibited absolute DIF values greater than 0.6 logits with four of them even exceeding differences of 1 logit. Thus, as a considerable number of items showed large DIF indicating great differences in the measurement of 9th graders (in the main study) and adults (in the link study). The same test seems to assess a different construct in the different populations. Note, that here the main study and link study differ not only by a large age difference (9th graders aged 16 to adults of age 21 to 78), but also in educational and occupational setting (school vs. mainly work), test setting (group testing in Grade 9 vs. individual testing at home for Adults), and most probably also in competence level. These differences between the populations seem to challenge the coherence of measurement.

Subsequent Analyses

In subsequent analyses we investigated the impact of item and test characteristics on the amount of DIF. Regarding item characteristics, we investigated whether DIF is related to item difficulty, text functions, cognitive requirements, and response format. We found no considerable relationship between DIF and text functions, cognitive requirements, or response format. Concerning text functions, the mean absolute DIF across all items and studies ranged from 0.24 (SD across studies [SDacross] being 0.08 and average SD within studies [SDwithin] being 0.18) for literary texts to 0.32 (SDacross=0.18, SDwithin=0.23) for commenting texts. Regarding cognitive requirements mean absolute DIF values across all items and studies were 0.33 (SDacross=0.13, SDwithin=0.26) for finding information in the text, 0.23 (SDacross=0.10, SDwithin=0.19) for drawing text-related conclusions, and 0.28 (SDacross=0.11, SDwithin=0.20) for reflecting and assessing. Complex MC items were affected with slightly lower absolute DIF values (M=0.23, SDacross=0.08, SDwithin=0.20), than MA items (M=0.28, SDacross=0.13, SDwithin=0.19) and simple MC items (M=0.29, SDacross=0.13, SDwithin=0.22). The impact of

the different text functions, cognitive requirements, and response formats was moderate and very similar for the different studies.

There was a strong relationship of DIF with item difficulty. Table 3 shows the correlation of item difficulty with both, absolute DIF value and DIF value. Note that DIF was calculated as the differences in estimated item difficulty in the link study minus the estimated item difficulty in the main study. Thus, positive DIF values indicate that an item is easier in the main study than in the link study and negative values that the item is easier in the link study than in the main study.

Table 3. Correlation of the value (DIF) and the absolute value (DIFabs) of differential item functioning and item difficulty (β) across studies and tests

Link	Test	*cor(β,DIFabs)*	*cor(β,DIF)*
G5-G7	G5	0.27	-0.26
	G7 easy	0.31	0.23
	G7 difficult	-0.48	-0.21
G7-G9	G7 easy	-0.33	0.30
	G7 difficult	0.25	0.05
	G9	-0.27	-0.25
G9-AD	G9	-0.03	-0.60
	AD	-0.00	-0.26

The correlation of item difficulty with absolute DIF indicates for which items DIF occurs; when positive, DIF tends to occur in difficult items; when negative, DIF tends to occur in easy items; and when zero, it tends to occur in both easy and difficult items. The sign of the correlation of item difficulty with DIF indicates in which direction DIF occurs. Positive values

indicate that easy items are easier in the link study than in the main study and difficult items are more difficult in the link study than in the main study. The results for the various studies (see Table 3) suggest a very heterogeneous picture with different correlation patterns for different studies. We had no theory on that and investigated the patterns exploratorily as possible explanations for DIF. We focus on the studies with large DIF, that is, the G7 easy and hard tests in the G7-G9 link and the G9 test in the G9-AD link. For the easy G7 test form in the link G7-G9, DIF mainly occurred for easy items (*cor(β, DIFabs)=-0.33*) with items being more difficult in the main study (on 7th graders) than in the link study (on 9th graders) (*cor(β, DIF)=0.30*). For difficult items, DIF hardly occured (*cor(β, DIFabs) =-0.33*). This is different for the respective difficult test form of the same study. For the G7 difficult test form, DIF mainly occured for difficult items (*cor(β, DIFabs)=0.25*). There was no relationship of item difficulty and the direction of DIF (*cor(β, DIF)=0.05*). Another pattern was found for the G9 test linking the G9 test to the adult test. Here DIF occured for easy and for difficult items (*cor(β, DIFabs) =-0.03*) with the easy items being more difficult in the link study (with an adult sample) and the difficult items being more difficult in the main study (with a sample of 9th graders) (*cor(β,DIF)=-0.60*). As the size and direction of DIF for different item difficulties varied a lot across studies, it is difficult to find an explanation. The results on measurement invariance for the Grade 7 easy test linking G7 to G9 seem to be affected by test targeting, with DIF occurring mainly for items with a low targeting (i.e., that are either too difficult or too easy for the respective sample). In fact, the Grade 7 easy test that was completed by 9th graders in the link study, yielded considerable ceiling effects for some items. About eight out of 29 items had a probability to be solved above 95%. When these items were excluded from the DIF analyses, the relative amount of DIF could be reduced. This is not necessarily true in the other studies, as in most studies (e.g., studies in Grade 9 and on Adults) item difficulty is rather low, but DIF occurs on items of all difficulties. Thus,

although there does not seem to be a clear pattern, there are indications that measurement variance is related to item difficulty and test targeting.

On test level, we evaluated whether similar results of measurement invariance can be found for assessing other competencies. This facilitates the drawing of conclusions concerning the extent to which the results depend on the specific test or are rather population specific. The main and link studies linking the Grade 9 test to the adult test were also used to link mathematical competence. For mathematical competence we found similar results on dimensionality and measurement invariance as for reading competence. There was hardly any DIF on the adult test, which was administered to the same population in the main study and the link study; there was, however, large DIF for items of the Grade 9 test (also see Pohl & Carstensen, 2013). Similar coherence across competence domains was found in the school cohorts, such as for linking ICT literacy from Grade 6 to Grade 9. In the Grade 9 ICT test that was administered to 9th graders in the main and the link study almost no DIF occurred (analogous to the results of reading competence linking Grade 7 to Grade 9). In contrast, some DIF was present in the Grade 6 test that was administered to 6th graders in the main study and to 9th graders in the link study with four out of 30 items exceeding DIF values of 0.6 logits (there was also DIF present in the respective analyses comparing measurement models between Grade 7 and Grade 9 on reading). In summary, different competencies that were assessed in the NEPS such as reading, mathematical competence, or ICT literacy showed similar patterns of measurement (in)variance across specific age spans.

Discussion

In the present study, we investigated whether it is possible to coherently measure reading competence across the life span within the NEPS. We specifically asked whether the reading

tests for different cohorts measure the same construct and whether each reading test measures the same construct across different samples. The results on dimensionality showed that within the same population tests for different age groups did measure the same construct. Thus, the tests were well constructed to assess the same construct coherently across the test forms. However, when the same test was administered to samples drawn from very different populations, the measurement models differed between samples, that is, measurement invariance did not fully hold. The more different the populations were, the larger DIF was found. Differences in populations are indicated by differences in age (e.g., linking Grade 5 and Grade 7), differences in educational and occupations settings (e.g., students in school and adults at work), differences in test settings (e.g., group testing in school, individual testing at home), and differences in competence levels (e.g., differences in the assignment to test forms in Grade 7). Only for linking Grade 5 to Grade 7, which were similar in educational setting and test setting, an adequate amount of measurement invariance could be assured. On test level, item difficulty and test targeting seem to play a role for results on measurement invariance.

The differences in item functioning for different populations may to some extent occur due to differences in test-taking behavior. This can, for example, be evaluated by missing values. While samples from similar populations in our study showed rather similar missing item patterns, samples from different populations differed in their missing item patterns. Adults in the main study and in the link study, for example, showed a very similar missing pattern on the amount of omitted and not reached items as well as non-valid responses. Students in Grade 9 and adults differed immensely in their missingness patterns. The adult sample reached fewer items, omitted more items, and produced more invalid responses than Grade 9 students. We also found greater correlations between the number of omissions and item difficulty for the student populations (correlations ranging from 0.23 to 0.55) than for adults (*cor*=0.12). The greater age and

competence level, the higher these correlations were in school. This may indicate that students in school, especially the older they are and the more competence they gained, apply a different test-taking strategy than adults. This is also corroborated by the finding that the number of omissions of adults is greater with lower competence levels (correlation of reading competence and number of omissions being -.26), while it is hardly related in the students samples (correlations ranging from -.07 for Grade 9 students to -.12 for Grade 5 students). Especially older and more competent students seem to use some test-taking strategy omitting difficult items. This fits well in the research on test wiseness (e.g., Diamond & Evans, 1972; Gibb, 1964; Millman, Bishop, & Ebel, 1965) and test motivation (e.g., Wise & DeMars, 2005, 2006), which also reports on omission of items and quitting on the test (e.g., Schmitt, Chan, Sacco, McFarland, & Jennings, 1999; Zerpa, Hachey, van Barnfield, & Simon, 2011). Investigating differences in test-taking behavior may help explain the results on differences in measurement invariance across different populations in further research.

There are some implications for large scale studies that can be drawn from our results. As our results show, although within the same sample, adjacent test forms may assess a unidimensional construct, the measurement model may differ for different populations. This is especially the case when differences between populations increase. Thus, for planning a longitudinal study that requires linking of test forms, the differences in the populations to be linked should be kept to a minimum. That means that linking should be performed across smaller age ranges. In NEPS, linking between Grade 9 and Adults did not prove successful, but possibly linking Grade 9 students to Grade 12 students and students in the school cohort to younger adults, might facilitate appropriate linking. Similarities in linked samples also include similarities in test settings. Mode effect studies may help assessing the effect of individual vs. group testing and, thus, accounting for it. This is done in NEPS in other age cohorts (Kröhne & Martens, 2011). As

it might be that DIF between different populations occurs due to differences in test-taking strategies, a more thorough instruction on how to take the test may help prevent from measurement variation. This issue can be approached in an even more sophisticated way, by computerized testing, where more control over item skipping and response time is possible.

In our study we focused on the prerequisites for linking. These results are very relevant for the NEPS, since they are the basis for choosing the actual linking models within the age cohorts and, if possible, across age cohorts as well. One of the outcomes of the NEPS will be an empirical answer to the question, whether and for which domains it is possible to construct a common scale across the life span. As far as the results presented here indicate, it will be feasible to construct common scales within some age limits.

In order to establish a common scale, one has to make assumptions about item drift. If one assumes that observed item drift is not due to any systematic reason like a shift in constructs, a link may be based on items that did not show DIF, assuming partial measurement invariance. One has to rely on the assumption that the items chosen for linking are not confounded by item drift. This, however, cannot be empirically tested. This assumption may be more plausible for the tests in the school cohorts, that is, linking Grade 5 to Grade 7 and Grade 7 to Grade 9. As the DIF on the Grade 9 test linking G9-AD is very large and the populations differ a lot, it may be less plausible here. Further link studies, e.g., linking G9 to G12, G12 to university students or tertiary students to younger adults may and will give more evidence to investigate whether linking across age cohorts will be possible.

After having evaluated the plausibility of different linking assumptions, the question is how to link different test forms. From research we know that different decisions in the scaling process typically lead to somewhat different vertical scales (Camilli et al., 1993; Loyd & Hoover, 1980; Williams, Pommerich, & Thissen, 1998; Yen, 1986). No consensus exists in the literature

as to which set of procedures produces the vertical scale that most adequately captures the nature of development (Kolen & Brennan 2004). It rather seems that the optimal linking model depends on the degree of violation of the assumptions made in a linking model given its particular design and sample sizes. In any way, within the NEPS different linking analyses, preferably linking with restrictions on item difficulty on an item level and as an alternative, linking with restrictions on item difficulty on the test level, will be explored to quantify the impact on the linking results. One of the crucial questions will be to decide which items are considered "undrifted" and thus will contribute to the link and which items are considered to show item drift and will be excluded from establishing the link. Consequently a thorough evaluation of the linking model applied to a particular study is needed. In order to quantify the degree of linkability, linking errors will be computed and compared. A possible solution for linking approaches for the NEPS might be to distinguish *strict linking* from linking of tests that might be considered as *connected* in a less stringent way. A strict link requires most items to be invariant over time resulting in small linking errors only, whereas connected tests may allow item drifts to occur more frequently and the link error might thus be larger. From a substantive NEPS point of view, to have connected test forms across different age cohorts may have the potential for relevant cohort comparisons in the NEPS, whereas following the competence development of students longitudinally over subsequent years will require a strict link assuming measurement invariance and small linking errors. The investigation of which linking models are appropriate in NEPS falls in the scope of further research.

Acknowledgement

This research used data from the National Educational Panel Study (NEPS). From 2008 to 2013, NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

This research is based on the dedicated work of professors and research assistants within the NEPS. We especially thank Karin Gehrer, Stefan Zimmermann, Cordula Artelt, and Sabine Weinert for developing the tests on reading competence, that are the basis of our research, and Maike Krannich, Michael Wenzler, Theresa Rohm, and Odin Jost for their valuable assistance in analyzing the data. Our thanks also go to the staff of the NEPS administration of surveys and to the methods group.

References

Alexandrowicz, R. (2008). Wieviel ist „ein bisserl“? Ein neuer Zugang zum BIC im Rahmen von Latent-Class-Analysen [How much is „a bit“? A new approach to the BIC within the framework of Latent Class Analyses]. In J. Reinecke & C. Tarnai (Eds.), *Klassifikationsanalysen in Theorie und Anwendung* (pp. 141-165). Münster: Waxmann.

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*, 5–14.

Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., et al. (2011). Sampling designs of the National Educational Panel Study: Challenges and solutions. *Zeitschrift für Erziehungswissenschaft, 14*, 51–65.

Blossfeld, H.-P., von Maurice, J., & Schneider, T. (2011). The National Educational Panel Study: Need, main features, and research potential. *Zeitschrift für Erziehungswissenschaft, 14*, 5–17.

Böhme, K., & Robitzsch, A. (2009). Methodische Aspekte der Erfassung der Lesekompetenz [Methodological aspects of reading assessment]. In D. Granzer, O. Köller, A. Bremerich-Vos, M., van den Heuvel-Panhuizen, K., Reiss, & G. Walther (Eds.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (pp. 250–289). Weinheim: Beltz.

Camilli, G., Yamamoto, K., & Wang, M. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement, 17*, 379–388.

Carstensen, C. H., Lankes, E. M., & Steffensky, M. (2012). Modellierung von längsschnittlichen Daten am Beispiel einer quasi-experimentellen Studie zur Erfassung von naturwissenschaftlichen Kompetenzen im Kindergartenalter [Modeling of longitudinal data illustrated on a quasi-experimental study of the assessment of scientific competencies in preschool children]. In W. Kempf & R. Langeheine (Eds.), *Item-Response-Modelle in der sozialwissenschaftlichen Forschung* (pp.109–126). Berlin: Regener Verlag.

Diamond, J. J., & Evans, W. J. (1972). An investigation of the cognitive correlates of testwiseness. *Journal of Educational Measurement, 9*, 145–150.

Doran, H. C., & Cohen, J. (2005). The confounding effect of linking bias on gains estimated from value-added models. In R. W. Lissitz (Ed.), *Value-added models in education: Theory and applications* (pp. 80–104). Maple Grove, MN: JAM Press.

Dorans, N. J., Pommerich, M., & Holland, P. (Eds.) (2007). *Linking and aligning scores and scales*. New York, NY: Springer.

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,* 50–79.

Gibb, B. G. (1964). *Testwiseness as secondary cue response* (Doctoral dissertation). Stanford University, Ann Arbor, Michigan: University Microfilms, 1964. No. 64-7643.

Haberkorn, K., Pohl, S., Carstensen, C., & Wiegand, E. (in press). Scoring of complex multiple choice items in NEPS competence tests. In H.-P. Blossfeld, J. von Maurice, M. Bayer, & J. Skopek (Eds.), Methodological issues in longitudinal surveys. Springer.

Haertel, E. (1991). *Report on TRP analyses of issues concerning within-age versus across-age scales for the National Assessment of Educational Progress*. Washington, DC: National Center for Education Statistics.

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

Hoover, H. D. (1984). The most appropriate scores for measuring educational development in the elementary schools: GE's. *Educational Measurement: Issues and Practice, 3*, 8–14.

Hoover, H. D., Dunbar, S. B., & Frisbie, D. A. (2003). *The Iowa Tests: Guide to research and development.* Chicago, IL: Riverside Publishing.

Jones, L. V. & Olkin, I. (Eds.). (2004). *The Nation's Report Card: Evolution and perspectives.* Bloomington, IN: Phi Delta Kappa Educational Foundation.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., et al. (Eds.) (2003). *The development of National Educational Standards. An expertise* (Vol. 1). Berlin: BMBF.

Kristen, C., Römmer, A., Müller, W., & Kalter, F. (2005). *Longitudinal studies for education reports – European and North American examples,* Report commissioned by the Federal Ministry of Education and Research. Bonn, Berlin: Federal Ministry of Education and Research (BMBF).

Kröhne, U. & Martens, T. (2011). Computer-based competence tests in the national educational panel study: The challenge of mode effects. *Zeitschrift für Erziehungswissenschaft, 14,* 169-186.

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

Linn, R. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6,* 83–102.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*, 179–193.

Martineau, J. (2006). Distorting value-added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Psychological Statistics, 31*, 35–62.

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

McClellan, C. A., Donoghue, J. R., Gladkova, L., & Xu, X. (2005). *Cross-grade scales in NAEP: Research and real-life experience.* Presentation at the conference Longitudinal Modeling of Student Achievement, Maryland Assessment Research Center for Education Success, University of Maryland, College Park, MD.

Millman, J., Bishop, D. H., & Ebel, R. (1965). An analysis of test wiseness. *Educational and Psychological Measurement, 25,* 707–726.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: ETS Policy Information Center.

Monseur, C., & Berezner, A. (2007). The computation of equating errors in international surveys in education. *Journal of Applied Measurement, 8,* 323–335.

Monseur, C., Sibberns, H., & Hastedt, D. (2008). Linking errors in trend estimation for international surveys in education. In M. von Davier & D. Hastedt (Eds.), *Issues and methodologies in large-scale assessments* (pp. 113–122). Hamburg: IEA-ETS Research Institute.

Mullis, I. V., Martin, M. O., Foy, P., & Drucker, K. T. (2012). *PIRLS 2011 international results in reading.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

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,* 80–109.

OECD (2013). *PISA 2012 Assessment and analytical framework: Mathematics, reading, science, problem solving, and financial literacy*. OECD Publishing.

Penfield, R. D. & Algina, J. (2006). A generalized DIF effect variance estimator for measuring unsigned differential test functioning in mixed format tests. *Journal of Educational Measurement, 43*, 295–312.

Pohl, S. (2014). Longitudinal multi-stage testing. *Journal of Educational Measurement, 50*, 447–468.

Pohl, S., & Carstensen, C. H. (2012). *NEPS Technical Report: Scaling the data of the competence tests* (NEPS Working Paper No. 14). Bamberg, Germany: University of Bamberg, National Educational Panel Study.

Pohl, S., & Carstensen, C. H. (2013). Scaling the competence tests in the National Educational Panel Study—Many questions, some answers, and further challenges. *Journal of Educational Research Online, 5*, 189–216.

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not reached items in competence tests-Evaluating different approaches accounting for missing responses in IRT models. *Educational and Psychological Measurement, 74,* 423–452.

Pollack, J. M., Atkins-Burnett, S., Najarian, M., & Rock, D.A. (2005). *Early Childhood Longitudinal Study, Kindergarten class of 1998–99 (ECLS–K), Psychometric report for the fifth grade.* U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Popham, W. J. (2000). *Educational measurement.* Boston, MA: Allyn and Bacon.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielsen & Lydiche.

Reckase, M. D. (2009). *Multidimensional item response theory.* New York, NY: Springer.

Reckase, M. D., & Martineau, J. A. (2004). *Growth as a multidimensional process.* Paper presented at the Annual Meeting of the Society for Multivariate Experimental Psychology, Naples, FL.

Robitzsch, A., Dörfler, T., Pfost M., & Artelt, C. (2011). Die Bedeutung der Itemauswahl und der Modellwahl für die längsschnittliche Erfassung von Kompetenzen: Lesekompetenzentwicklung in der Primarstufe [Relevance of item selection and model selection for assessing the development of competencies: The development in reading competence in primary school students]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 43,* 213–227.

Rock, D. A., Pollack, J. M., Owings, J., & Hafner, A. (1990). *Psychometric report for the NELS:88 base year test battery.* Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Rock, D. A., Pollack, J. M., & Quinn, P. (1995). *Psychometric report of the NELS: 88 base year through second follow-up.* Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Rupp, A. A. & Vock, M. (2007). National educational standards in Germany: Methodological challenges for developing and calibrating standards-based tests. In D. Waddington, P. Nentwig, & S. Schanze (Eds.), *Making it comparable: Standards in science education* (pp. 173–198). Münster: Waxmann.

Schmitt, N., Chan, D., Sacco, J. M., McFarland, L. A., & Jennings, D. (1999). Correlates of person fit and effect of person fit on test validity. *Applied Psychological Measurement, 23,* 41–54.

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*, 139–161.

Thissen, D. (2012). *Validity issues involved in cross-grade statements about NAEP results.* Washington, DC: American Institutes for Research, NAEP Validity Studies Panel.

Tong, Y., & Kolen, M. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education, 20*, 227–253.

von Davier, A. A., Carstensen, C. H., & von Davier, M. (2008). Linking competencies in horizontal, vertical and longitudinal settings and measuring growth. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 121–149). New York: Hogrefe & Huber.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York: Springer-Verlag.

von Maurice, J., Artelt, C., Blossfeld, H.-P., Faust, G., Rossbach, H.-G., & Weinert, S. (2007). *Bildungsprozesse, Kompetenzentwicklung und Formation von Selektionsentscheidungen im Vor- und Grundschulalter: Überblick über die Erhebungen in den Längsschnitten BiKS-3-8 und BiKS-8-12 in den ersten beiden Projektjahren* [Educational processes, competence development and formation of selection decisions in preschool and primary school age: An overview of the first two years of data collection in the longitudinal studies BiKS-3-8 and BiKS-8-12]. Bamberg: Otto-Friedrich-Universität.

Wang, S., & Jiao, H. (2009). Construct equivalence across grades in a vertical scale for a K-12 large-scale reading assessment. *Educational and Psychological Measurement*, *69*, 760–777.

Wang, S., Jiao, H., & Zahng, L. (2013). Validation of longitudinal achievement constructs of vertically scaled computerized adaptive tests: A multiple-indicator, latent-growth modeling approach. *International Journal of Quantitative Research in Education, 1,* 383–407.

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen, C. H. (2011). Development of competencies across the life span. *Zeitschrift für Erziehungswissenschaft, 14*, 67–86.

Williams, V. S. L., Pommerich, M., & Thissen, D. (1998). A comparison of developmental scales based on Thurstone methods and item response theory. *Journal of Educational Measurement, 35*, 93–107.

Wise, S. L. & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10,* 1–17.

Wise, S. L. & DeMars, C. E. (2006). An application of item response time: The effort moderated model. *Journal of Educational Measurement, 43,* 19–38.

Wu, M. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement, 29*, 15–27.

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

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23*, 299–325.

Zerpa, C., Hachey, K., van Barnfield, C., & Simon, M. (2011). Modeling student motivation and students' ability estimates from a large-scale assessment of mathematics. *SAGE Open, 1,* 1–9.

3.3 Manuscript 3: Aggregation of Complex Multiple Choice Items

Original Source of Publication

Haberkorn, K., Pohl, S., Carstensen, C., & Wiegand, E. (in press). *Scoring of complex multiple choice items in NEPS competence tests.* In H.-P. Blossfeld, J. von Maurice, M. Bayer, & J. Skopek (Eds.). *Methodological issues in longitudinal surveys.* Springer.

Copyright

Scoring of Complex Multiple Choice Items in NEPS Competence Tests

Kerstin Haberkorn,[1] Steffi Pohl,[2] Claus Carstensen,[1] Elena Wiegand[3]

Abstract

In order to precisely assess the cognitive achievement and abilities of students, different types of items are often used in competence tests. In the National Educational Panel Study (NEPS), test instruments also consist of items with different response formats, mainly simple multiple choice (MC) items in which one answer out of four is correct and complex multiple choice (CMC) items comprising several dichotomous "yes/no" subtasks. The different subtasks of CMC items are usually aggregated to a polytomous variable and analyzed via a partial credit model. When developing an appropriate scaling model for the NEPS competence tests, different questions arose concerning the response formats in the partial credit model. Two relevant issues were how the response categories of polytomous CMC variables should be scored in the scaling model and how the different item formats should be weighted. In order to examine which aggregation of item response categories and which item format weighting best models the two response formats of CMC and MC items, different procedures of aggregating response categories and weighting item formats were analyzed in the NEPS, and the appropriateness of these procedures to model the data was evaluated using certain item fit and test fit indices. Results suggest that a differentiated scoring without an aggregation of categories of CMC items best discriminates between persons. Additionally, for the NEPS competence data, an item format weighting of one point for MC items and half a point for each subtask of CMC items yields the best item fit for both MC and CMC items. In this paper, we summarize important results of the research on the implementation of different response formats conducted in the NEPS.

Keywords: item response theory, partial credit model, complex multiple choice, response category aggregation, item format weighting, scoring

[1] University of Bamberg, Germany.
[2] Free University Berlin, Germany.
[3] University of Mannheim, Germany.

Scoring of Complex Multiple Choice Items in NEPS Competence Tests

1. Item Formats and Scaling Model of the NEPS Competence Tests

In the process of test development, the choice of the items' format plays a crucial role for different aspects of validity (Rodriguez, 2002). So far, comprehensive item writing rules and guidelines have been published (Downing & Haladyna, 2006; Haladyna & Rodriguez, 2013; Osterlind, 1998), and a variety of analyses have been performed on different item formats in order to evaluate the strengths and weaknesses of each response format. A main distinction is usually made between selected response (SR) items and constructed response (CR) items. Whereas constructed response items require the examinee to create a response to a specific question or item stem, selected response items require choosing an answer out of a set of options or matching options to several stems that are presented. Most assessments make use of the SR item format (Osterlind, 1998). SR items ensure an efficient and effective measurement, and a large body of research shows that thoroughly and representatively constructed SR items achieve high content validity (Downing, 2006; Haladyna & Downing, 2004; Rodriguez, 2002). Furthermore, the objective, efficient scoring prevents threats to validity, such as construct-irrelevant variance induced by the subjectivity of human raters (Haladyna & Rodriguez, 2013).

In the National Educational Panel Study (NEPS), different types of SR items are used in the competence tests. In the NEPS, the tests measuring mathematical competence, reading competence, scientific literacy, and information and communication technologies (ICT) literacy mainly include simple multiple choice (MC) and complex multiple choice (CMC) items1 (see Pohl & Carstensen, 2012, for a more detailed description of the different response formats; for an overview of the competencies, see also Weinert et al., 2011). MC items in the NEPS usually consist of four response options, with one being correct and three being incorrect. CMC items in the NEPS are composed of a number of subtasks, with one out of two response options being correct. An example for an MC and a CMC item is presented in Figure 1. The number of subtasks within CMC items varies in the NEPS competence tests.

[1] Note that some test instruments in the NEPS additionally contain matching items as another type of SR item and constructed response items, but these response formats are rare and thus not considered in the analyses.

Mr. Brown owns a rectangular piece of land and wants to fence it in. He has already made some calculations and then bought a 40 *m* fence. The piece of land has a width of 8 *m*. How long is the land?

☐	5 *m*
☐	8 *m*
☐	12 *m*
☐	16 *m*

(a)

Are the following statements about the study's result correct?

	yes	no
Half of the participants showed at least one side effect, because 50 is half of 100.	☐	☐
Sickness occurred less than itching, because 50+40 is less than 50+70.	☐	☐
About 53% of the participants showed at least one side effect, because (50+40+70)/3 ≈ 53%.	☐	☐
More than half of the participants showing sickness also showed itching, because 50:90 > 50%.	☐	☐

(b)

Figure 1. Example of (a) an MC item and (b) a CMC item within NEPS competence tests (Neumann et al., 2013).

As CMC items consist of item bundles with a common stimulus, the assumption of local item independence may be violated within CMC items (e.g., Yen, 1993). To account for this local item dependence (LID), the subtasks within CMC items are usually aggregated to polytomous super-items, as suggested by many researchers (e.g., Andrich, 1985; Ferrara, Huynh, & Michaels, 1999). Several psychometric models have been developed for polytomous variables. The item bundles may, for example, be analyzed via a graded response or a partial credit model (Huynh, 1994; Wainer, Sireci, & Thissen, 1991). For scaling the NEPS competence data, a partial credit model (Masters, 1982) was used. The partial credit model was deliberately chosen because of its

membership in the family of Rasch models and the advantageous properties that Rasch models are known to have (Penfield, Myers, & Wolfe, 2008). For scaling the competence data, many large-scale studies, for example, PISA or NEPS, use one-parameter (1PL) models or extensions of this model to preserve the item weights intended by the instrument construction (see Pohl & Carstensen, 2012, for an argumentation of model choice in the NEPS). If the number of items from different conceptual aspects is intentionally chosen, the 1PL scaling model ensures the intended weightings of the conceptual aspects in contrast to the 2PL model, in which the items' weight depends on their empirical factor loadings. Given the 1PL model, we asked ourselves how we could best implement the different response formats in the scaling model and especially how we should score the categories of the CMC items and how we should weight both MC and CMC items.

2. Research on the Implementation of Response Formats Within a Scaling Model

Until now, several methods of implementing items with different response formats in a 1PL-scaling model have been applied in large-scale studies. The scoring procedures for items with different response formats, in particular, differed in their degree of aggregation of categories they used for polytomous variables as well as in their weighting of the item formats. In the following section, first, common aggregation approaches for response categories of CMC items are presented, and second, weightings of different item formats within an Item Response Theory (IRT) framework are described.

2.1 Aggregation

The simple MC items are usually scored dichotomously, with one point given for a correct response and zero points given for the selection of an incorrect response (also called distractor). Reviewing various competence assessments that implemented different response formats, there are two widely applied aggregation methods for polytomous variables. First, the *All-or-Nothing scoring rule* is very common and means that subjects only receive full credit if all answers on subtasks are correct (Ben-Simon, Budescu, & Nevo, 1997). If at least one subtask is answered incorrectly, the person receives no credit. This method makes use of a dichotomous scoring and is implemented for CMC items in the study "Teacher Education and Development Study in Mathematics" (TEDS-M, see Blömeke, Kaiser, & Lehmann, 2010). Another established method of dealing with CMC items is the *Number Correct (NC) scoring rule*, which rewards partial

knowledge, meaning that partial credit is given for each correctly solved subtask of a CMC item (see Ben-Simon et al., 1997). To apply the NC scoring rule, the subtasks of CMC items are formed to a composite score, and each of the categories receives partial credit according to the number of correctly answered subtasks. This scoring option is well known and has often been used in large-scale studies, such as PISA (Adams & Wu, 2002).

While several researchers have examined the impact of the two aggregation options for CMC items using parameters of classical test theory (CTT), there are only few results within the field of IRT. Hence, findings of research based on CTT are described first to get an impression of the impact of the two aggregation options before presenting results based on IRT. Based on CTT-analyses, Ben-Simon and colleagues (1997) reported a disadvantage of the All-or-Nothing scoring rule for students with low ability since the students' partial knowledge is not captured. They pointed out that the NC scoring, in particular, measures lower-performing students more accurately. Hsu (1984) and Wongwiwatthananukit, Bennett, and Popovich (2000) demonstrated advantages of the NC scoring rule regarding reliability and discrimination. Nevertheless, Hsu found only a slight increase in discrimination and reliability of the NC scoring in comparison with the All-or-Nothing scoring rule and thus argued that the slight gains of the NC scoring do not seem to justify the additional effort involved in this procedure in comparison with dichotomous scoring.

Si (2002) compared the effects of NC scoring and dichotomous scoring using IRT. In his study, he applied several dichotomous and polytomous IRT-models to simulated item-response data and investigated effects on parameter estimation using different model parameterizations (1-, 2-, and 3PL) and degrees of aggregation (dichotomous versus polytomous). His results provided evidence that polytomous models produce more accurate ability estimates than dichotomous models independent of the prior distribution of the persons' abilities. Furthermore, the 1PL model considerably outperformed the 2PL- and 3PL models. Among the polytomous models, the partial credit model exhibited the most accurate ability estimation. Nevertheless, Si only examined the effect of various models on the accuracy of the estimated person abilities.

2.2 Weighting of Different Response Formats

Besides their variation in the degree of aggregation of response categories within polytomous CMC items, competence assessments also differ in their allocation of scores for solving items

with different response formats. PISA, for instance, awards one point for correctly solved MC items. The CMC items are given different maximum scores based on theoretical considerations by the test developers (OECD, 2009). There are a few CMC items with special requirements that are therefore scored with a maximum score of two points. Other CMC items are weighted equally to the simple MC items and are hence given a maximum score of one point when all subtasks are solved correctly. During the development of scaling models for the NEPS competence data, the question arose of whether CMC items should receive the same maximum score as simple MC items or whether they should have more impact on the overall competence score. One may argue that CMC items should be scored equally to MC items to make sure that the different items in the test contribute equally to the competence score. Others may suggest that CMC items should be weighted more as they incorporate a set of tasks and each subtask should get the same maximum score as an MC item. CMC items contain two response options, whereas simple multiple choice items consist of four response options. Thus, an appropriate procedure might also be a scoring of half points for each subtask while MC items receive one point when solved correctly.

Up to now, there has been only little research on weighting different types of item formats, especially concerning the item formats implemented in the NEPS competence tests. In contrast, differential weighting of items has received considerable attention in scaling test instruments. In the field of CTT, different methods and principles for weighting items have been established (Ben-Simon et al., 1997; Kline, 2005; Stucky, 2009). Overall, the weighting of items is usually performed using a statistical or theoretical approach. If item weighting is based on statistical data, items' reliability and factor loadings may be regarded. Weighting items by objective theoretical criteria involves weighting determined by experts or weights imposed by items' length, difficulty, or assumed validity. In the field of IRT, studies mainly focused on models with an implicit item weighting in 2- or 3-PL-models (Stucky, 2009). However, studies dealing with a priori weighting of response formats in IRT models to preserve the item weighting by construction are limited. Lukhele and Sireci (1995) as well as Sykes and Hou (2003) looked for ways to model different response formats with deliberately chosen weights via IRT. Lukhele and Sireci established a specific weighting of MC and constructed response (CR) items in a 1PL-model using "unweighted" IRT marginal reliabilities for weighting the different formats. Sykes and Hou also applied a priori weighting of MC and CR items to their test data by giving a maximum score of one point for each MC item and a maximum score of two points for each CR item, but they did

not examine different weighting schemes to find out the best way to implement the response formats. In sum, these studies used a priori weighting for implementing response formats in an IRT framework, but fit indices of the response formats were not evaluated as important indicators for the appropriateness of the weighting procedure. Furthermore, only constructed response items and simple MC items were implemented, whereas CMC items, which are included in the NEPS competence data, were not.

Given the limited findings on the implementation of response formats in a 1PL model, different analyses were conducted in the NEPS in order to replicate and extend preliminary research into the best way to deliberately model different item formats. Two relevant questions concerning the response formats in the development of the scaling model that were addressed in the NEPS were as follows: *First*, to which degree should the response categories of CMC items be aggregated, and *second*, how should the response formats encompassing CMC and MC items be weighted assuming that both item types assess the same latent trait?

In the following section, we begin by illustrating the empirical study we carried out to find the best aggregation option for the CMC items in the NEPS. Second, we describe the NEPS research of Haberkorn, Pohl, and Carstensen (2015), who looked for the best weighting procedure of different response formats for the NEPS competence tests.

3. Investigating Aggregation for CMC Items in NEPS Competence Tests

3.1 Method

Sample and Instruments

For analyzing the impact of different aggregation schemes for CMC items in the scaling model, data from two competence domains, which were assessed in a main study of ninth graders in the National Educational Panel Study, were used. In the main study in Grade 9, the subjects were engaged in different competence tests. The analyses were conducted using the domains of *scientific competence* and *information and communication technologies (ICT) literacy*. The tests of scientific competence assessed children's scientific knowledge in the contexts of health, environment, and technology (Hahn et al., 2013). The ICT instrument tapped children's ability to locate and use essential information and their knowledge on different kinds of technology, such

as hardware and software (Senkbeil, Ihme, & Wittwer, 2012). The competence tests of scientific competence and ICT literacy contained a reasonable amount of MC and CMC items (see Schöps & Saß, 2013; Senkbeil & Ihme, 2012).

Since cases with less than three valid responses were excluded from the IRT analyses, the analyses were undertaken based on 14,301 subjects for scientific competence and 14,312 subjects for ICT literacy.[1] The test instrument to assess scientific competence consisted of 19 simple MC items and nine CMC items. The number of subtasks within the CMC items varied from four to six items. The test instrument of ICT literacy included 32 MC items and eight CMC items, and there were four to seven subtasks within the CMC items.

Analyses

The partial credit model (Masters, 1982) was used to apply the different scoring approaches to the data. Marginal maximum likelihood estimation was chosen for estimating the models, and all analyses were done using ConQuest (Wu, Adams, Wilson, & Haldane, 2007). If at least one of the subtasks of CMC items contained a missing value, the whole CMC item was coded as missing response. According to Gräfe (2012) as well as Pohl, Gräfe, and Rose (2013), ignoring missing responses in the scaling model yields unbiased item- and person parameter estimates. Therefore, missing responses were ignored in the application of the different scoring procedures. If response categories of the polytomous CMC items had less than 200 cases, adjacent categories were combined to avoid possible estimation problems. This occurred for the lowest categories, in particular, and predominantly if the CMC item consisted of many subtasks. For scientific competence, the two lowest categories of a CMC variable were collapsed into one category and received a score of zero points within four CMC items. For ICT literacy, the lowest categories of zero and one were combined into one category within seven CMC items due to low cell frequencies.

Different aggregation schemes for the categories of polytomous items were applied to the data. The MC items were always scored as zero points for an incorrect answer and as one point for a correct answer. In order to examine the impact of aggregation of response categories, CMC items were scored a) dichotomously, with one point given if all subtasks were answered correctly and

[1] Note that due to later updates and data-editing processes, the number of persons and items may slightly differ from the number of persons and items found in the Scientific Use File.

zero points otherwise. This resembles the All-or-Nothing scoring rule implemented for most of the CMC items in PISA. In contrast, the second rule b) was a more differentiated scoring according to the NC scoring rule, with a maximum score of one point for a correct response on all subtasks and partial credit for each correctly answered subtask. The partial credit points ranged between zero points and one point in equal intervals. As a consequence, the partial credit steps were different depending on the number of categories within the CMC item. For example, the categories of a CMC item with five categories were scored with a score of r = 0, 0.25, 0.5, 0.75, and 1, whereas the categories of a CMC item with four categories were scored r = 0, 0.33, 0.67, and 1.

To get detailed information about changes in item- and test parameters caused by the two aggregation options, the CMC items were first analyzed separately without considering MC items, and different item statistics were investigated. We evaluated difficulty, correlation of the item score of CMC items with the total score (discrimination value as computed in ConQuest), and test reliability of the two aggregation rules. The correlation of the item score with the total score corresponds to the product-moment-correlation between the categories of CMC items and the total score, and the correlation is labeled as discrimination in the following sections. Furthermore, based on analyses of both MC and CMC items, the range of the abilities of test takers with partially correct answers was explored in order to assess the amount of information that is lost by applying a dichotomous scoring. For this purpose, differences between person ability in the second-highest and the lowest response categories were computed for each polytomous item. For example, for a CMC item with 4 subtasks, subjects with only incorrect answers might have a medium ability of -0.54 logits (the estimate of person ability in each category is always computed using the other items in the test only), whereas subjects who solved three out of the four subtasks might have a medium ability of 0.03 logits. Thus, person ability between the lowest and the second-highest response category in this case would vary with a range of 0.57 logits. This range of person ability is combined into one category in the All-or-Nothing scoring rule. Therefore, a computation of the range of person abilities is performed to investigate how much information we lose if we analyze these persons together in one category.

3.2 Results

First, we present the comparison of the two aggregation procedures for the categories of CMC items, the All-or-Nothing scoring, and the NC scoring. In Table 2, the item difficulty and discrimination for the All-or-Nothing scoring and the NC scoring in the Science and ICT domains are depicted.

Table 1

Item Location Parameters, Characterizing the Items' Difficulty (in Logits), and Discrimination of the All-or-Nothing Scoring and the NC Scoring

	Science				ICT			
	Location parameter		Discrimination		Location parameter		Discrimination	
	All-or-Nothing scoring	NC scoring	All-or-Nothing scoring	NC scoring	All-or-Nothing scoring	NC scoring	All-or-Nothing scoring	NC scoring
CMC_1	-0.30	-4.11	0.47	0.48	0.38	-2.57	0.50	0.53
CMC_2	1.58	-1.34	0.41	0.49	0.73	-3.63	0.50	0.49
CMC_3	1.02	-3.39	0.46	0.45	0.79	-2.02	0.45	0.42
CMC_4	0.33	-2.47	0.57	0.56	0.61	-3.47	0.56	0.56
CMC_5	0.26	-3.17	0.57	0.58	0.46	-2.73	0.48	0.50
CMC_6	-0.24	-2.39	0.52	0.56	0.24	-2.93	0.57	0.59
CMC_7	0.92	-2.58	0.55	0.54	2.01	-2.16	0.44	0.62
CMC_8	0.02	-2.34	0.50	0.54	1.75	-1.20	0.36	0.50
CMC_9	0.63	-2.48	0.55	0.58	For ICT, there were only 8 CMC items.			
Means	0.47	-2.70	0.51	0.53	0.87	-2.59	0.48	0.53

Note. The analyses for these results were undertaken using CMC items only.

With regard to item difficulty, high differences between the All-or-Nothing scoring and the NC scoring emerged. The NC scoring for CMC items yielded considerably lower difficulty estimates than the All-or-Nothing scoring. Comparing the two aggregation options by the average item difficulties, their means differed by about 3.17 logits (standard deviation (*SD*) = 0.71) for Science and 3.46 logits (*SD* = 0.69) for ICT. Thus, substantially higher item difficulties were estimated for the All-or-Nothing scoring than for the NC scoring since subjects with partially correct answers were given no credit in the All-or-Nothing scoring and there were consequently more subjects with zero points on the items. Furthermore, the item discrimination varied slightly to moderately between the dichotomous scoring and the NC scoring. For most of the items in

Science and ICT, discrimination at the item level increased when applying the NC scoring. For six out of the 17 items, rather equal discriminations occurred. Overall, the average discrimination showed moderate gains resulting in more differentiated measures for the NC scoring.

Differences between the two aggregation options were even more evident when comparing the reliability. For the Science domain, the NC scoring (EAP/PV reliability = 0.652, WLE reliability = 0.595) yielded higher reliability estimates than the All-or-Nothing scoring (EAP/PV reliability = 0.593, WLE reliability = 0.433). The reliability improved substantially for the NC scoring (EAP/PV reliability = 0.518, WLE reliability = 0.444) (especially for ICT) in comparison with the All-or-Nothing scoring (EAP/PV reliability = 0.444, WLE reliability = 0.150).

In order to evaluate the possible loss of information in the application of the All-or-Nothing scoring, the range of the abilities of persons within the categories that were collapsed in the dichotomous scoring was examined. For a reliable estimation of these abilities, the analyses were performed based on MC and CMC items. The range of person abilities for each CMC item was computed as the difference between the medium ability of subjects who were in the second-highest category and the medium ability of subjects in the lowest category (see Table 2).

Table 2

Range of the Abilities (in Logits) of Persons Who Answered Incorrectly or Only Partially Correctly

	Science		ICT	
Item	Number of categories	Range of abilities	Number of categories	Range of abilities
CMC_1	3	0.83	3	0.67
CMC_2	3	0.72	4	0.86
CMC_3	4	0.82	5	-0.16
CMC_4	5	0.51	5	0.47
CMC_5	4	1.00	3	0.80
CMC_6	3	0.47	5	0.74
CMC_7	4	0.57	6	1.02
CMC_8	4	0.79	4	1.00
CMC_9	4	0.90	--	--

For example, regarding the first CMC item of the ICT test, which contained three categories, the range of person abilities within the base to the second categories was 0.67 logits, indicating that subjects reaching the second category had a higher overall ability by 0.67 logits on average than subjects who didn't solve any of the subtasks of the CMC item. In the dichotomous scoring, these categories within CMC items (for Item 1 in ICT category 0-2) were collapsed and scored with zero points.

For Science, the test consisted of nine CMC items, and persons who received no or only partial credit varied substantially in their general ability (computed across the other items in the test), with M = 0.73 logits (SD = 0.18) on average. The highest differences occurred for Item 5. Subjects who solved three out of the four subtasks correctly had a higher overall ability by about one logit than subjects who didn't solve any subtasks correctly for this item. However, the persons who differed considerably in their ability were treated equally in the NC scoring. Eight CMC items were included in the ICT test, and persons who were collapsed into one group in the dichotomous scoring also exhibited substantial variation in their overall estimated ability (M = 0.68, SD = 0.38), except for Item 3. This item had an unsatisfactory item fit, and the persons who didn't solve any of the subtasks correctly had a higher ability by 0.16 logits than persons who solved four fifths of the subtasks of the CMC item. In this case, the reversed range of abilities underlines the misfit of the item to the model.[1] Overall, the analyses of the abilities' range indicate that persons who received no or only partial credit differed greatly in their general ability.

Taking together the impact of the two aggregation options on item difficulty, discrimination, test reliability, and person's range of abilities with no or partially correct answers, the results provide evidence for rather high gains in information about subjects' competencies using the NC scoring instead of the All-or-Nothing scoring.

4. Overview of Research on Weighting of Response Formats in NEPS Competence Tests

The question of how to appropriately weight different NEPS response formats in a 1PL model was investigated in an elaborate study by Haberkorn et al. (2015), and the main findings of the study are presented in the following section. In order to examine the impact of different

[1] Due to unsatisfactory item fit, this item was not included in the Scientific Use File.

weighting schemes of CMC and MC items on the item parameters, Haberkorn et al. made analyses based on the same NEPS competence data of Science and ICT from the main study in G9 which was used for exploring the influence of aggregating CMC items. Since items with low item fit statistics were excluded from the final dataset (Schöps & Sass, 2013; Senkbeil & Ihme, 2012), the analyses of weighting were based on 9 CMC and 19 MC items in Science as well as 10 CMC and 17 MC items in ICT. Three different weighting procedures were compared by Haberkorn and her colleagues, and for each of the options, the categories of the CMC items were given partial credit. As a consequence, the degree of aggregation did not differ among the different weighting options. This allowed for disentangling item weighting from the aggregation procedure for the response categories. The implemented weighting options were as follows: The correctly solved MC items were always scored with one point. The CMC items a) were given a maximum score of one point to equal their weight to the MC items, b) were scored by giving half points per category to reflect the reduced number of two response options within the subtasks instead of four response options in the MC items, and c) received one point per category, and the subtasks of the CMC items were thus weighted equally to the simple MC items. An example of the different scoring options used for a CMC item is depicted in Table 3.

Table 3

Example for Different Scoring Methods of a CMC Item With Six Categories

	Three weighting options		
Categories of a CMC item with five subtasks	(a) Maximum score is 1	(b) Half points per correct subtask	(c) One point per correct subtask
0	0	0	0
1	0.2	0.5	1
2	0.4	1	2
3	0.6	1.5	3
4	0.8	2	4
5	1	2.5	5

Haberkorn et al. (2015) compared the weighted mean square (WMNSQ) and the respective *t*-value of the three scoring options in order to investigate the best a priori weighting for the two

response formats of CMC and MC items. It is important to note that Haberkorn et al. used different statistical parameters for the evaluation of the weighting of item formats than for the evaluation of different aggregation options depending on the amount of information the parameters provided. The aggregation procedures, in particular, differed in their reliability and discrimination estimates but did not differ much in their WMNSQ estimates. The different weighting options also had different discrimination estimates, but the WMNSQ and corresponding *t*-value were more appropriate for an evaluation of the weighting options in order to find the most balanced fit for MC and CMC items within the Rasch model.

First, we present the main results for the Science domain found by Haberkorn et al. (2015). The impacts of the three weighting procedures for CMC items in relation to MC items (which were always scored with one point for a correct answer) are depicted in Figures 2 and 3: an equal weighting of MC and CMC items with a maximum score of one point, half points per subtask of CMC items, or one point per subtask for CMC items. Figure 2 includes means and standard deviations of the WMNSQ, separately computed across MC and CMC items, for the three different scoring options. Figure 3 depicts means and standard deviations of the *t*-value for the three different scoring options, separately computed across MC and CMC items.

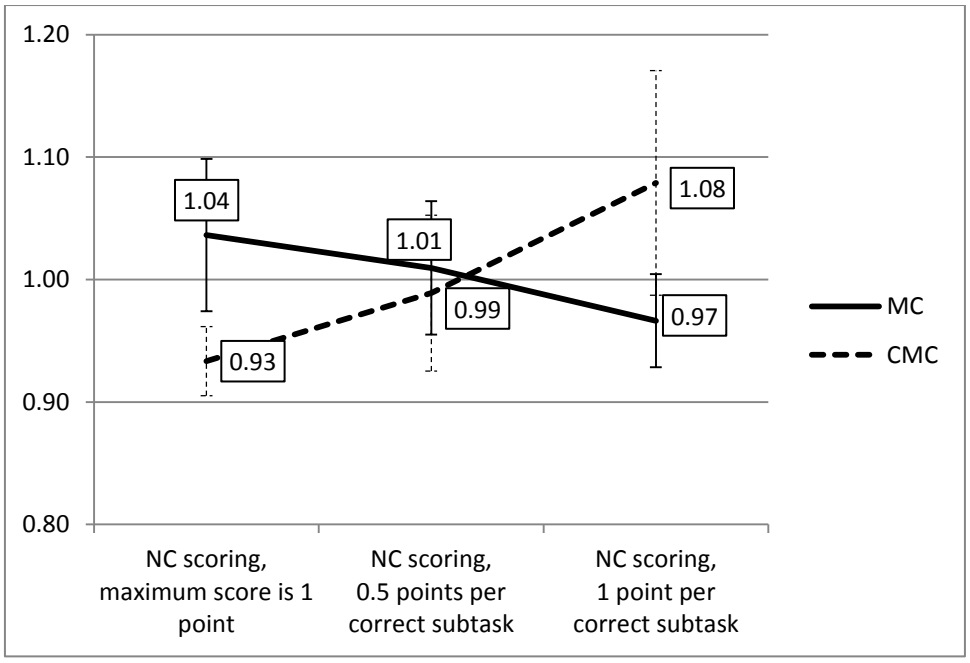


Figure 2. Means and standard deviations of the WMNSQ for different item weightings in the domain of Science (Haberkorn et al., 2015).

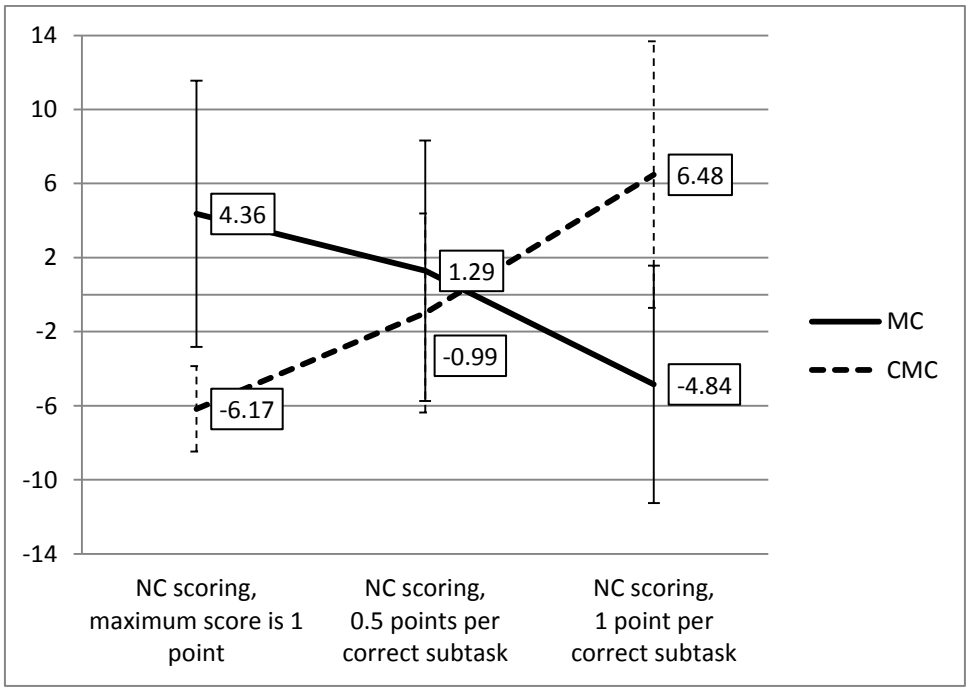


Figure 3. Means and standard deviations of the *t*-value of the WMNSQ for different item weightings in the domain of Science (Haberkorn et al., 2015).

As can be seen in these figures, an equal weighting of MC and CMC items, which meant that MC items as well as the polytomous CMC items were scored with a maximum of one point, resulted in an underfit for MC items and an overfit for CMC items. Both the WMNSQ (see Figure 2) and, more evident due to the rather large sample size, the *t*-value of the WMNSQ (see Figure 3) indicated that MC as well as CMC items did not fit the underlying model well . In contrast, the opposite was found to be true when each of the subtasks of CMC items was weighted equally to MC items and when correct responses to MC items as well as correctly solved subtasks of CMC items were consequently given one point in the scaling model. In this case, an overfit of MC items and a rather large underfit of CMC items emerged. A scoring of half points per category for the CMC items yielded the best item fit for the WMNSQ and the respective *t*-value. When the categories of the CMC items were given half of the weight of MC items, both MC and CMC items showed the most balanced fit.

Haberkorn et al. (2015) applied the same weighting procedures of CMC items in relation to MC items to the ICT data (see Table 4).

Table 4

Means and Standard Deviations (in Parentheses) of the WMNSQ and Corresponding t-Values for the Three Weighting Options in the Domain of ICT Literacy (Haberkorn et al., 2015)

Response format	Fit criterion	NC scoring, maximum score is 1	NC scoring, half points per correct subtask	NC scoring, one point per correct subtask
MC items	WMNSQ	1.02 (0.06)	1.00 (0.06)	0.97 (0.05)
	t-value	1.66 (6.75)	-0.06 (6.90)	-4.51 (6.87)
CMC items	WMNSQ	0.93 (0.04)	0.99 (0.03)	1.15 (0.05)
	t-value	-6.21 (3.30)	-0.26 (2.02)	11.41 (4.53)

Note. Correctly solved MC items were always scored with one point.

When looking at the WMNSQ and the respective *t*-value, the results of Science were replicated. An equal weighting of the MC items and the CMC items consisting of several subtasks caused an overfit of CMC items and a slight underfit of MC items. Conversely, with an equal weighting of the subtasks of CMC items to MC items, the CMC items showed a large underfit, and the MC items showed a slight overfit. Taking the fit of MC and CMC items together, the best fit of the weighted items to the model was given when each of the categories of CMC items was scored with half points. While a scoring of half points per category still resulted in a slight underfit of MC items in the Science domain, the same scoring option caused a quite optimal fit for both MC and CMC items for ICT (Haberkorn et al., 2015).

Haberkorn et al. (2015) also applied a restricted 2PL model in which loadings within response formats were set equal but were allowed to vary between response formats. By regarding the two discrimination indices for MC and CMC items, they received the empirical weight of the response formats. As expected, the values were close to 0.5. In addition to applying the different weighting approaches to NEPS competence data, Haberkorn et al. studied the impact of the weighting options on fit indices in PISA competence tests. Their results replicated the findings of the NEPS research and demonstrated that weighting the subtasks of CMC items with half of the weight of MC items yielded a quite appropriate fit of MC and CMC items to the model.

5. Conclusion and Discussion

The aim of this chapter was to provide an overview of major research issues concerning the implementation of MC and CMC items in a Rasch model addressed in the NEPS. According to

often-applied scoring procedures in competence assessments and based on theoretical deliberations, the impact of different degrees of aggregating response categories within polytomous CMC items was explored in the NEPS, and the appropriateness of different weighting schemes was investigated.

With regard to the aggregation options, the comparison of the All-or-Nothing scoring and the Number Correct scoring showed clear evidence of the discriminating effect of the NC scoring. To avoid a loss of information, CMC items should be scored as differentiated as possible. The application of a dichotomous scoring for CMC items may implicate the assumption that subjects answering no subtask correctly and subjects answering some subtasks of an item correctly do not differ in their ability. Indeed, the current investigation has documented that there is considerable variation in ability within these subjects. Thus, following the suggestions of other researchers (Si, 2002), NC scoring should be preferred over All-or-Nothing scoring to improve the accuracy of ability estimates. However, limitations in the application of NC scoring may arise due to low cell frequencies in certain categories. In this case, categories within CMC items may be collapsed in the scaling of the data in order to avoid estimation problems (OECD, 2009; Pohl & Carstensen, 2012, 2013).

The investigation of different weighting schemes for CMC items in relation to MC items carried out by Haberkorn et al. (2015) pointed consistently to the fact that a scoring of about half a point for the categories within CMC items while awarding one point per MC item matches the empirical data quite well. In contrast, the other weighting procedures performed substantially worse in the Science and ICT domains. Of course, the relative weight of MC and CMC items might differ with regard to other age groups, competence domains, or large-scale studies. Competence assessments that aim at assessing other abilities and skills using these item formats might obtain other suitable scoring schemes. In the development of a 1PL scaling model, it therefore seems crucial to empirically evaluate weights that are constituted theoretically a priori. As argued by Haberkorn et al. (2015), a combination of applying 2PL models in the development of a scaling model and using a priori weights in the final application of a 1PL model may hence serve as a promising procedure for competence assessments to implement theoretically constituted features and, simultaneously, enhance the statistical properties of the scaling model.

The analyses computed by Haberkorn et al. included the main item formats within NEPS competence tests; recommendations for weighting item formats are thus restricted to CMC and MC items. Further research on response formats applied in other large-scale studies, such as constructed response items, will be useful to extend weighting guidelines. Finally, studies on competence tests in other age groups, competence domains, and national as well as international studies will be of interest to expand upon the current understanding of the best way to comprise different response formats in a scaling model.

Acknowledgement

This paper uses data from the National Educational Panel Study (NEPS): Sarting Cohort Grade 9, doi:10.5157/NEPS:SC4:4.0.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

6. References

Adams, R., & Wu, M. (2002). *PISA 2000 technical report*. Paris, France: OECD.

Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. In N. Brandon-Tuma (Ed.), *Sociological methodology* (pp. 33–80). San Francisco, CA: Jossey-Bass.

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.

Blömeke, S., Kaiser, G., & Lehmann, R. (2010). *TEDS-M 2008 – Professionelle Kompetenz und Lerngelegenheiten angehender Primarstufenlehrkräfte im internationalen Vergleich.* Münster, Germany: Waxmann.

Downing, S. M., & Haladyna, T. M. (Eds.). (2006). *Handbook of test development*. Mahwah, NJ: L. Erlbaum.

Downing, S. M. (2006). Selected-response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3–26). Mahwah, NJ: Erlbaum.

Ferrara, S., Huynh, H., & Michaels, H. (1999). Contextual explanations of local dependence in item clusters in a large-scale hands-on science performance assessment. *Journal of Educational Measurement, 36*(1), 119–140.

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

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

Haberkorn, K., Pohl, S., & Carstensen, C.(2015). *Incorporating different response formats of competence tests in an IRT-model.* Manuscript in preparation.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17–27.

Haladyna, T. M., & Rodriguez, M. C. (2013) *Developing and validating test items*. New York, NY: Routledge.

Hsu, T. C. (1984). The merits of multiple-answer items as evaluated by using six scoring formulas. *Journal of Experimental Education*, *52*(3), 152–158.

Huynh, H. (1994). On equivalence between a partial credit item and a set of independent Rasch binary items. *Psychometrika, 59,* 111-–119.

Kline, T. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks, CA: Sage.

Lukhele, R., & Sireci, S. G. (1995, April). *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.

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

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

OECD (2009*). PISA 2006 technical report*, Paris, France: OECD.

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

Osterlind, S.J. (1998*) Constructing Test Items: Multiple-Choice, Constructed-Response, Performance, and Other Formats.* Dordrecht, Netherlands: Kluwer Academic.

Penfield, R. D., Myers, N. D, & Wolfe, E. W. (2008). Methods for assessing item, step, and threshold invariance. Polytomous items following the partial credit model. *Educational and Psychological Measurement*, *68*(5), 717–733.

Pohl, S., & Carstensen, C. H. (2012). *NEPS technical report – Scaling the data of the competence tests.* (NEPS Working Paper No. 14). Bamberg: University of Bamberg, National Educational Panel Study.

Pohl, S., & Carstensen, C. H. (2013). Scaling the competence tests in the National Educational Panel Study – Many questions, some answers, and further challenges. *Journal of Educational Research Online, 5*, 189–216.

Pohl, S., Gräfe, L., & Rose, N. (2013). Dealing with omitted and not reached items in competence tests – Evaluating approaches accounting for missing responses in IRT models. *Educational and Psychological Measurement, 74,* 423–452.

Rodriguez, M. (2002). Choosing an item format. In G. Tindal, & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 213–231). Mahwah, NJ: Erlbaum.

Schöps K., & Saß, S. (2013). *NEPS technical report for science – Scaling results of Starting Cohort 4 in ninth grade*. (NEPS Working Paper No 23). Bamberg: University of Bamberg, National Educational Panel Study.

Senkbeil, M. & Ihme, J. M. (2012). *NEPS technical report for computer literacy – Scaling results of Starting Cohort 4 in ninth grade* (NEPS Working Paper No. 17). Bamberg: University of Bamberg, National Educational Panel Study.

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 of Educational Research Online, 5,* 139–161.

Si, C. B. (2002). *Ability estimation under different item parameterization and scoring models* (Doctoral dissertation). Retrieved from http://digital.library.unt.edu/ark:/67531/metadc3116/m2/1/high_res_d/dissertation.pdf

Stucky, B. D. (2009). *Item response theory for weighted summed scores* (Master's thesis). Retrieved from https://cdr.lib.unc.edu/indexablecontent?id=uuid:03c49891-0701-47b8-af13-9c1e5b60d52d&ds=DATA_FILE

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

Wainer, H., Sireci, S. G., & Thissen, D. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*(3), 237–247.

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. von Maurice (Eds.), *Education as a lifelong process: The German National Educational Panel Study (NEPS) (pp. 67-86).* Wiesbaden: VS Verlag für Sozialwissenschaften.

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.

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. (2007). *ACER ConQuest 2.0 – Generalised item response modelling software*. Camberwell, Australia: ACER Press.

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

3.4 Manuscript 4: Dimensionality and Weighting of Multiple Choice and Complex Multiple Choice Items

Original Source of Publication

Haberkorn, K., Pohl, S., & Carstensen, C. (2015). Incorporating different response formats of competence tests in an IRT model. Manuscript submitted for publication.

Incorporating Different Response Formats of Competence Tests in an IRT Model

Kerstin Haberkorn,[1] *Steffi Pohl,*[2] *Claus Carstensen*[1]

Abstract

Competence tests within large-scale assessments usually contain various task formats to adequately measure the participants' knowledge and skills. Two response formats that are frequently used are simple multiple choice (MC) items and complex multiple choice (CMC) items. When incorporating these response formats in a scaling model, they are mostly assumed to be unidimensional. In empirical studies different empirical and theoretical schemes of weighting CMC items in relation to MC items have been applied to construct the overall competence score. However, the dimensionality of the two response formats and the different weighting schemes have only rarely been evaluated. The present study, thus, addressed two questions of particular importance when implementing MC and CMC items in a scaling model: Do the different response formats form a unidimensional construct and, if so, which of the weighting schemes considered for MC and CMC items appropriately models the empirical competence data? Using data of the National Educational Panel Study, we analyzed scientific literacy tests embedding MC and CMC response formats. We cross-validated the findings on another competence domain and on data of another large-scale assessment. The analyses revealed that in all competence domains and studies the different response formats form a unidimensional measure. Additionally, we found evidence that the a priori weighting scheme of giving each subtask of CMC items half the points of an MC item models the response formats' impact on the competence score quite appropriately. Implications of the findings for the development of scaling models including different response formats are discussed.

Keywords: item response theory, complex multiple choice, item format weighting, scoring, dimensionality

[1] University of Bamberg, Germany.
[2] Free University Berlin, Germany.

Incorporating Different Response Formats of Competence Tests in an IRT Model

International large-scale assessments as well as national studies on students' achievement have to deal with the challenge of efficiently and precisely measuring different competencies of the participants. When operationalizing theoretical constructs of the competencies to be measured, one relevant issue refers to the choice of the items' format. To increase strengths and compensate weaknesses of each format, Martinez (1999) recommended a combination of item formats in test instruments. Taking validity and variation into account, competence tests in (large-scale) assessments, for example the Program for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS), the National Assessment of Educational Progress (NAEP), or the National Educational Panel Study (NEPS), hence usually contain different response formats to comprehensively assess the subjects' competencies (Allen, Donoghue, & Schoeps, 2001; OECD, 2012; Olson, Martin, & Mullis, 2008).

A common classification of item formats is the differentiation between selected-response (SR) and constructed-response (CR) formats (Haladyna & Rodriguez, 2013; Osterlind, 1998). SR items consist of correct and incorrect options to a problem and require the examinee to select one or several options. In CR items no options are presented, but the examinee has to generate the answer usually by writing down a word or short sentences. McMillan (2000) outlined that in comparison to CR formats such as essays, oral questions, or observations, SR items have the broadest spectrum in measuring competencies and skills. As SR formats are the most widely used item types in achievement tests of large-scale studies (Bleske-Recheck, Zeug, & Webb, 2007; Osterlind, 1998), in the following we focus on the common SR formats.

The two most well-established types of SR items in competence tests are multiple choice items and true-false items (Osterlind, 1998). The well-known multiple choice (MC) item encompasses an item stem, that is a question or an incomplete sentence, and different choices of responses, most conveniently four or five options comprising the correct answer and wrong answers, the so-called distractors (Haladyna & Rodriguez, 2013). True-false items are a popular variation of the MC format and require the examinee to make a binary choice (Haladyna, 1992). Often, true-false items are arranged to complex multiple choice (CMC) items that include a number of "true/false" statements. CMC items are, for instance, applied, in the PISA or NEPS study (Adams & Wu, 2002; Pohl & Carstensen, 2013). Note that the term complex multiple choice item is not used consistently in the literature. In recent large-scale studies such as PISA or NEPS it denotes multiple true-false items, while other researchers used the term slightly different for MC items with response options in which combinations of correct answers are offered (e.g., Haladyna & Rodriguez, 2013; Scalise & Gifford, 2006). In the following, we refer to CMC items as items including several binary subtasks as a synonym to multiple true-false items.

So far, large-scale studies have varied in their incorporation of MC and CMC item response formats for scaling the competence data. However, there is only little research on how the two response formats can be treated adequately in a scaling model. Specific questions that arise when implementing the response formats in a scaling model are: Do MC and CMC items measure the same latent trait? What impact should MC and CMC items have on the overall competence score? Should they be weighted equally in the scaling model? Should CMC items with more subtasks have a larger impact on the overall competence score? The purpose of the present study was to approach these questions by compiling theoretical considerations about the response formats and by thoroughly analyzing empirical data. Through a systematic investigation

of the questions concerning dimensionality and weighting on a variety of competence tests we aimed at delineating implications for implementing the two response formats in a measurement model.

Dimensionality of MC and CMC Items

In the following, we start by theoretically describing the cognitive processes accompanied with the response formats. We outline similarities and differences of MC and CMC items that might be of relevance for the question of whether the two response formats form distinguishable subdimensions. We then review empirical research on dimensionality of the two response formats.

Cognitive processes associated with MC and CMC items. As different item formats may activate different cognitive processes, several authors have highlighted the importance of considering the mental operations involved in answering items of different response formats (Haladyna & Rodriguez, 2013; Martinez, 1993, 1999; Palmer & Devitt, 2007; Snow, 1993). Scalise and Gifford (2006) proposed a comprehensive taxonomy of item formats and arranged many classic and innovative types of items according to the dimensions *constraint* and *complexity* (see Figure 1). They described relevant features of the item types, ranging from most constrained to least constrained response formats. In the most constrained item types, that is, the fully selected response formats, all components for the answer are supplied in advance. In the least constrained item types, that is, the fully constructed response formats, examinees are required to show complex performances such as projects, portfolios, or experiments without format constraints. Additionally, within each step of constraint, Scalise and Gifford sorted item formats by increasing complexity. As it is difficult to compare complexity between different

degrees of constraint, they were especially concerned with the constraint dimension of the response formats.

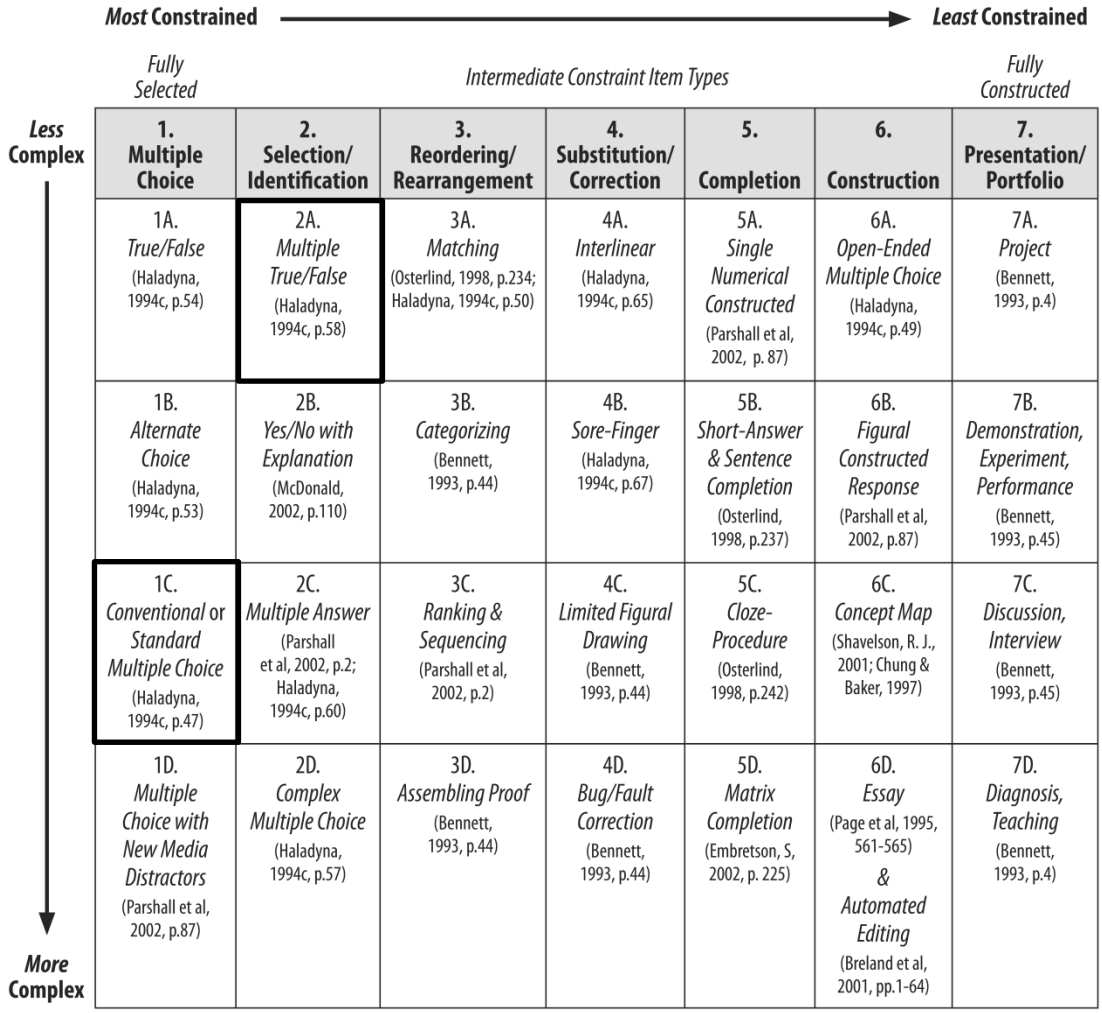


Figure 1. Classification system for different response formats. The response formats are arranged related to their constraint and their complexity. Adapted from "Computer-based assessment in E-learning. A framework for constructing 'intermediate constraint' questions and tasks for technology platforms" by K. Scalise and B. Gifford, 2006, *Journal of Technology, Learning, and Assessment, 4,* p. 9. Copyright 2006 by the Journal of Technology, Learning, and Assessment. Reprinted with permission.

The taxonomy shows that the two well-known SR-formats (1C. and 2A. in the figure) are located quite closely regarding their degree of constraint. The conventional MC item is the more restricted one of the two as it requires the subject to choose only one answer from a set of

response options. Van den Bergh (1990) analyzed the intellectual processes associated with MC items of a reading comprehension test based on Guilfords Structure-of-Intellect model (1971) and found that processes of recall, namely divergent and convergent production, as well as processes of recognition, namely cognition and evaluation abilities, are involved in solving the MC tasks. He did not find any differences in the cognitive abilities involved in MC and CR items. Rather, the participants differed individually regarding their particular intellectual abilities involved. Some of the participants, for instance, used evaluation strategies when solving the reading comprehension items while others did not. Other studies gave evidence that MC items can assess both lower-level thinking, such as recall of knowledge, as well as complex cognitions, such as evaluation or problem solving across content and grade (Coderre, Harasym, Mandin, & Fick, 2004; Haladyna, 1997; Haladyna, 2004; Hamilton, Nussbaum & Snow, 1997).

The multiple true-false item format is placed near the MC format with regard to its constraint. In the multiple true-false item format the choices within the item increase and so the degree of constraint decreases. In contrast to conventional MC items, the true-false items demand the subject to mentally generate a counterexample of the response option, because the two response alternatives of a true-false item are not explicitly proposed (Haladyna & Rodriguez, 2013). Some researchers critized the large guessing component of true-false items (Grosse & Wright, 1985; Haladyna & Downing, 1989), others stressed the benefits in testing time and test reliability (Frisbie, 1992; Ebel, 1970; Ebel & Frisbie, 1991). Haladyna (1992) pointed out that CMC items are well suited to measure low-level as well as higher-level skills.

Comparing the two response formats, similarities of the MC and the CMC format arise from the similar degree of constraint since both formats ask for answering questions or statements by making choices out of a set of options. Accordingly, both formats require on the

one hand to activate prior knowledge and process it and, simultaneously, evaluate different options. Differences might result from the different number of options that have to be evaluated and from the kind of options that are either presented directly or have to be created mentally. A series of studies showed that lots of MC items have only one or two well-functioning distractors so the number of options actually considered in MC items might be lower than the number of options presented (Haladyna & Downing, 1993; Lord, 1977; Rodriguez, 2005). Another difference might result from the dependence among subtasks in CMC items. Because of the same item stem and the close connection of response options one option might cue another one (Yen, 1993). However, dependencies among multiple true-false items seem not to be large (Albanese & Sabers, 1988; Frisbie & Druva, 1986). Finally, differences in item functioning might be induced by format familiarity, because performance on items increases with increasing familiarity of item formats (Fuchs et al., 2000).

In conclusion, from a psychological point of view it seems likely that MC and CMC items are quite similar concerning their mental processes yielding no additional sources for multidimensionality. After comparing the main cognitive facets associated with the two SR formats, the following section deals with results of empirical studies on the dimensionality of such response formats.

Research on dimensionality of mixed-format tests. In educational assessments, MC and CMC items are usually scaled using unidimensional models (e.g., OECD, 2012; Pohl & Carstensen, 2012). So far, dimensionality of items with different response formats has mainly been investigated for SR and CR items. Yet, little is known about whether the assumption of unidimensionality in tests including MC and CMC items holds in empirical studies.

Thus, we begin by reviewing research on MC and CR item formats and try to draw conclusions from the findings on MC and CMC item response formats. Overall, there are ambivalent results on dimensionality of SR and CR formats across different studies. Some researchers reported on multidimensionality in tests with SR and CR formats (Ackerman & Smith, 1988; Birenbaum & Tatsuoka, 1987; Ward, Frederiksen, & Carlson, 1980). Birenbaum and Tatsuoka (1987), for instance, administered SR and CR items assessing arithmetic abilities to students. Their analyses revealed that both tests had a different structure. Other researchers hold opposing views stating that MC and CR items are measuring quite the same latent traits (Bacon, 2003; Hohensinn & Kubinger, 2012; Thissen, Wainer, & Wang, 1994). In a meta-analysis Rodriguez (2002, 2003) explored the comparability of SR and CR item formats with variations in item stem and content. Even when the items were not stem-equivalent, but the content to be measured was intended to be the same, correlations were quite high. Traub (1993) investigated whether MC and CR items measured the same construct using data from different domains. He showed that MC and CR items were quite congeneric with respect to the abilities measured for reading comprehension and other quantitative domains, whereas in the writing domain the different item formats formed a multidimensional structure. For the science domain, Manhart (1996) also reported on multidimensionality based on item formats. For the domain of computer science, Bennett and his colleagues (1990) found evidence for unidimensionality. In sum, results on dimensionality of MC and CR items are somewhat equivocal. Because MC and CR items differ more in terms of their constraint (see Figure 1) than MC and CMC items, we assumed that studies on the dimensionality of MC and CMC items might provide less mixed results.

Overall, there are few studies that investigated dimensionality of CMC and MC items. Downing, Baranowski, Grosso, & Norcini (1995) included CMC items as well as MC items in a

medical achievement test in order to examine dimensionality. Their analyses exhibited that the two tests, that were intended to assess the same content, were highly correlated with latent correlations varying between 0.89 and 0.97. However, regarding the criterion-related validity, the MC items were higher correlated to an external performance variable than the CMC items. Using a test for second language ability, Dudley (2006) explored concurrent validity of MC and CMC items. The latent correlations between the variables formed by the two response formats ranged between .64 and 1.00 in vocabulary and reading, depending on the test form.

Altogether, results on dimensionality concerning the two SR formats are limited and not fully consistent. Nevertheless, information on dimensionality is crucial, as a unidimensional scale score might lead to biased parameter estimates, when the response formats form empirically distinguishable components (Walker & Beretvas, 2003). One focus of our study was, hence, to examine dimensionality of MC and CMC item response formats in different empirical competence data.

Weighting of MC and CMC Items in the Scaling Model

Assuming that the different response formats measure the same latent trait, the question of the relative weight of each item for constructing the overall competence score is raised. Reviewing weighting procedures for competence tests with mixed formats, we found that the studies differ considerably in their allocation of scores for the different response formats.

When researchers develop their scaling model for mixed-format competence data, the MC items are commonly scored dichotomously with one point awarded for the correct answer and zero for choosing one of the distractors. Before scoring CMC items, their subtasks are usually aggregated to polytomous super-items to account for local item dependence, as suggested by

many researchers (e.g., Andrich, 1985; Ferrara, Huynh, & Michaels, 1999). Subsequently, the polytomous items are given (partial) credit scores depending on the number of correctly solved subtasks. The scores assigned for the different response formats vary across different studies. In the following, the two main approaches in weighting different item formats are presented. Overall, item weighting may be determined empirically or may be based on theoretical deliberations (Kline, 2005; Ben-Simon, Budescu, & Nevo, 1997; Stucky, 2009).

Empirical weighting of different response formats. If an implicit empirical item weighting is chosen, the items' reliability, factor loadings, item-to-total correlation coefficients, or testing time may be used for determining item weights. Recently, the latent trait approach using IRT modeling has become a rather popular alternative to the traditional factor analytic approach. Some IRT models, for instance, the two-parameter (2PL) or three-parameter (3PL) logistic model allow for a simultaneous calibration of the different item types and for individual weights for each item as a function of the relation between the item and the underlying construct (e.g., Rutkowski, von Davier, & Rutkowski, 2013). In the 2PL model (or the 3PL model) a discrimination parameter for each item is estimated in addition to a location parameter (and a guessing parameter in case of the 3PL model) giving optimal empirical weights to the items. Large-scale studies such as the TIMSS or the IGLU study use 2- or 3PL models with an empirical item weighting based on statistical grounds. During calibration, the models assign more weight to items that -from a statistical perspective- carry more information for the underlying construct. Consequently, different types of items may be given different weights in the calibration depending on their discrimination. Hence, the 2- or 3PL model enables to statistically model the empirical item characteristic curves more closely, resulting in a better fit of the measurement model to the data compared to a 1PL model. However, as the empirical discrimination is allowed

to vary across all items, the relative weights within one item type and, hence, the contribution to the overall score may differ as well. A disadvantage of these IRT models might, thus, be that theoretical aspects such as an equal weighting of different subfacets of the construct, or an equal weighting of items with the same response format cannot be implemented in the scaling model. Hence, the final score does depend on statistical properties of the items, not on theoretical deliberations about the composition of the trait estimate.

A priori weighting of different response formats. Many large-scale studies, for example PISA or NEPS, do not use 2- or 3PL models, but use the one parameter (1PL) model or extensions of this model for scaling the data. In 1PL models the weight of the items is modeled only by the a priori scoring of the responses, as no additional discrimination parameter is estimated. As a consequence, an advantage of the 1PL model is that it preserves the item weights intended with the test construction and, thus, facilitates a theoretically driven development of the scaling model (see, for instance, Pohl & Carstensen, 2012, for an argumentation of model choice in NEPS). A popular model for dichotomous and polytomous items assessing competence domains in the family of Rasch models is the partial credit model (PCM; Masters, 1982). It is applied in PISA as well as in NEPS for mixed-format tests. When applying the PCM model to a mixed-format test, the weights for the different response formats are explicitly chosen before item calibration. Usually, these weights are assigned based on theoretical considerations (e.g., OECD, 2009).

Ercikan et al. (1998) specified different ways to explicitly weight diverse response formats. Two common a priori weighting schemes are a) equal weights for different item types, or b) weighting according to the complexity of an item or the number of subtasks of an item. With regard to the two SR item types, the first scoring rule implies awarding one point per MC item and per CMC item. Consequently, the MC items are weighted equally to the CMC items

independent of the number of subtasks in the CMC item. The second scoring rule means that one point per MC item is awarded and as many points for a CMC item as it contains subtasks.

In PISA, the choice of the scoring is based on theoretical deliberations of the test developers (OECD, 2009). Correctly answered MC items are given one point. Some of the CMC items are scored with a maximum of two points to reflect the special requirements in the particular tasks, while most of them are scored with a maximum of one point (equal to the MC items). In the Teacher Education and Development Study in Mathematics, the CMC items are scored with one point, if all subtasks are answered correctly (Blömeke, Kaiser, & Lehmann, 2010). Thus, the CMC items are weighted equally to MC items. In NEPS, the test developers determined that the subtasks of CMC items are given half the weight of an MC item. They want to reflect the fact that a subtask of a CMC item encompasses half the number of response options of an MC item. As only two response options have to be evaluated, only about half the amount of recall, recognition, and evaluation processes are required in CMC items. So, each correct answer to a subitem is awarded with half a point in the NEPS, whereas a correct answer to an MC item is awarded with one point.

Up to now, there have been no studies examining how well the different a priori weighting schemes resemble empirical competence data. Empirical results of the weighting schemes might therefore enable to evaluate the different a priori weighting schemes and explore how adequately they reflect the amount of information carried by the item response formats.

Research Questions

Already Osterlind (1998) warned about combining item formats incautiously when creating a common scale, as the interpretability of the scores may be suspect and even spurious.

One challenge for tests including mixed response formats may be multidimensionality of the different response formats. Applying unidimensional models to multidimensional data might bias the empirical parameter estimates and reduce the score precision. Whereas a lot of research has been undertaken to study dimensionality of CR and SR item formats, there is still a lack of evidence for different types of SR item formats. A comparison of the involved cognitive processes of MC and CMC items and first empirical results indicated that the two common SR response formats might assess the same latent trait. To verify this hypothesis, we empirically examined whether MC and CMC items served as an additional source for multidimensionality.

Assuming unidimensionality of the response formats, the question arises of how to weight different response formats within the scaling model. We, thus, aimed to investigate how well different a priori weighting schemes fit the empirical competence data. On the basis of the weighting rules by Ercikan et al. (1998) as well as weighting rules that have been applied in other large-scale studies, we specifically examined three a priori weighting schemes: a) CMC and MC items receive the same maximum score, b) each subtask of a CMC item receives the same maximum score as an MC item, and c) a scoring of half points for each subtask of a CMC item. Furthermore, we compared the results of the a priori weighting rules with an empirical weighting. Finally, we investigated whether the results can be generalized across contents and studies.

Method

Design and Sample

We addressed the research questions using data from the NEPS (Blossfeld, Roßbach, & von Maurice, 2011; Blossfeld, von Maurice, & Schneider, 2011). The NEPS aims at tracking students' developmental progress across the life span and, in particular, at measuring the

evolvement of competencies, conditions for their acquisition, and interactions with other variables. Measures tapping domain-general and domain-specific cognitive competencies as well as meta-competencies are implemented in the assessment (Weinert et al., 2011). The large-scale study comprises six main samples including newborns, Kindergarten children, secondary school children (fifth grade and ninth grade), students, and adults (Aßmann et al., 2011). These starting cohorts were first assessed between 2009 and 2012 and are now followed up longitudinally in order to obtain a broad data basis for analyzing educational processes. The subjects are surveyed yearly, competence tests are administered at larger intervals. All the participants in the starting cohorts are representatively sampled from German inhabitants.

Data from two scientific literacy tests of the NEPS were used for the analyses, as the scientific literacy tests embodied a substantial amount of CMC items in addition to MC items. One of the tests was administered in 2010 in Grade 9, and the other test was administered in Grade 6 in 2012. We chose two different grades in order to explore the research questions of our study in students of different ages. Cases with less than three valid responses were excluded from the analyses, because no reliable person ability score could be estimated for these students. Note that the number of subjects in the analyses presented in this paper and in the Scientific Use File may slightly differ due to data cleaning issues in the NEPS. In the analyses of the scientific literacy test in Grade 9, n = 14.301 students were included, 50.0 % of them were female, the students were on average M_{age} = 15.01 $(SD_{age}$ = 0.63) years old, and 94.1 % of them were born in Germany. In Grade 6, data of n = 4.871 students were used for the analyses and 48.5 % of them were female. The sample was on average M_{age} = 11.93 $(SD_{age}$ = 0.49) years old and 96.1 % of them declared Germany as country of birth.

To evaluate whether the results may be generalized, we cross-validated our findings in other studies and on other domains. For the cross-validation on a different competence domain, we employed data of an ICT competence test of the NEPS, that was administered in 2010 to 9th graders. Having excluded subjects with less than three valid answers, the final data set contained n = 14.485 subjects with 49.8 % being female. The students had an average age of M_{age} = 15.01 *(*SD_{age} = 0.63) years and 90.5 % of them were born in Germany.

For the cross-validation of the results in another large scale study, we drew on data of the Programme for International Student Assessment (PISA) study. PISA is a large international comparative study of achievement measuring performance of children aged 15 in about 70 countries by now (OECD, 2009, 2012, 2014). The survey was first conducted in 2000 and is now repeated every 3 years with competence assessments in reading, math, and science. The most recent data of scientific literacy assessed in nearly 70 countries in 2012 was used for the analyses to validate the results of the NEPS tests (OECD, 2013, 2014). We, again, used the scientific literacy test data, because this test in PISA featured the highest amount of MC and CMC items in comparison to the test instruments of the other domains. Again, cases with less than three valid answers were removed from the analyses. Altogether, n = 331.821 subjects entered the analyses, 50.5 % of them were female. The students were on average M_{age} = 15.78 *(*SD_{age} = 0.29) years old. In sum, 91.0 % of them were born in the country in which they took the competence test.

Measures and Procedures

The different competence tests in the NEPS primarily consist of MC and CMC item formats. An example for an MC and a CMC item in the NEPS tests is depicted in Figure 2.

Mr. Brown owns a rectangular piece of land and wants to fence it in. He has already made some calculations and then bought a 40 *m* fence. The piece of land has a width of 8 *m*. How long is the land?

☐	5 *m*
☐	8 *m*
☐	12 *m*
☐	16 *m*

(a)

Are the following statements about the study's result correct?

	yes	no
Half of the participants showed at least one side effect, because 50 is half of 100.	☐	☐
Sickness occurred less than itching, because 50+40 is less than 50+70.	☐	☐
About 53% of the participants showed at least one side effect, because (50+40+70)/3 ≈ 53%.	☐	☐
More than half of the participants showing sickness also showed itching, because 50:90 > 50%.	☐	☐

(b)

Figure 2. Example for (a) an MC item and (b) a CMC item in the NEPS competence tests (Neumann et al., 2013).

MC items in NEPS usually consist of four response options with one being correct and three being incorrect. CMC items in NEPS are composed of a number of subtasks with one out of two response options being correct. The proportion of different types of SR item formats in the NEPS competence tests may be considered typical, as Osterlind (1998) pointed out that the most commonly used SR item formats are MC items followed by true-false items.

The instruments assessing scientific literacy in the NEPS are constructed based on an elaborated conceptual framework. They are intended to assess children's scientific knowledge in health, environment, and technology (Hahn et al., 2013; Schöps & Saß, 2013). The test on scientific literacy in Grade 9 comprises 28 items. 19 of them are simple multiple choice items with one answer out of four being correct. Nine of these items are complex multiple choice items in the form of multiple true-false items where the examinee has to decide at each option whether the answer is correct or not. The CMC items include three to six subtasks, most of them have four subtasks. The test on scientific literacy in Grade 6 consists of 27 items with 17 of them being simple MC items and 10 of them being CMC items. All CMC items contain four options in a true/false format.

The test on ICT literacy in the NEPS is constructed to measure different facets of technological and information literacy (Senkbeil & Ihme, 2012; Senkbeil, Ihme, & Wittwer, 2013). After dropping items with an unsatisfactory item fit, the ICT test in Grade 9 encompassed 36 items (Senkbeil & Ihme, 2012). Twenty nine items had an MC item response format, seven were presented in the CMC response format. The CMC items contained four to seven options in a true/false format, most of them had four or six options. The tests assessing scientific literacy and ICT in the NEPS were administered as paper-and-pencil tests in a group setting at school with a testing time of about 30 minutes per competence domain.

In PISA most of the items are MC items. Furthermore, the competence tests encompass CMC items and some CR item types. The science assessment in PISA requires students to identify scientific issues, to explain phenomena scientifically, and to use scientific evidence (OECD, 2013). As in the NEPS, items on knowledge of science and knowledge about science are implemented in the tests. The scientific literacy test consists of MC and CMC items as versions

of SR items, and CR items which may be coded automatically, rated by a manual, or rated by experts. Overall, the science assessment in 2012 incorporated 16 CMC items, 18 MC items, and 21 CR items. In the present study, MC and CMC items were retained in the analyses and CR items were excluded, because our study focused on MC and CMC response formats. The PISA tests were administered in paper-and-pencil format and the subjects had to complete tests of different domains in about two hours testing time (for additional information see OECD, 2013).

Analyses

All data were scaled using IRT. Missing responses were ignored in the parameter estimation (Gräfe, 2012; Pohl, Gräfe, & Rose, 2014). All specifications of 1PL models were made with ACER ConQuest (Wu, Adams, Wilson, & Haldane, 2007). The models referring to the 2PL family were estimated with the software mdltm (von Davier, 2005).

Dimensionality. In order to examine dimensionality of the competence tests, a unidimensional and a two-dimensional partial credit model were applied to the data of each of the four studies. In the two-dimensional model, which was specified as a between-item multidimensional random coefficients multinomial logit model (Adams, Wilson, & Wang, 1997), two latent variables were modeled. The MC items loaded on one latent dimension, the CMC items loaded on the other latent dimension. In the one-dimensional model, one latent variable was used for all items. The partial credit model from the family of Rasch models was chosen for the uni- and the two-dimensional model in accordance with the scaling procedure in NEPS (Pohl & Carstensen, 2012; 2013) and in PISA (OECD, 2014). For the analyses, the subtasks of each CMC item were aggregated to a polytomous variable. To avoid possible estimation problems, categories with less than 200 valid responses were subsumed with the adjacent category (Pohl & Carstensen, 2012). In accordance with the scoring in NEPS (Haberkorn, Pohl, Carstensen, & Wiegand, 2015; Pohl &

Carstensen, 2012), each category of the polytomous items was scored with half points. Thus, the examinees received partial credit for correctly solved subtasks of a CMC item. Different criteria were used for the evaluation of dimensionality. We particularly regarded the correlation between the latent variables formed by MC and CMC items. Additionally, we compared the unidimensional and the multidimensional model by using two overall fit indices from information theory: the Akaike Information Criterion (AIC; Akaike, 1974) and the Bayesian Information Criterion (BIC; Schwarz, 1978).

Weighting. Three common a priori scoring schemes were applied to each of the competence tests. As before, the partial credit model was used for the analyses. In all analyses the MC items were scored with one point if answered correctly, and zero points otherwise. The CMC items were formed to polytomous variables and partial credit was given according to the number of correctly answered subtasks. The scoring of the CMC items was varied systematically. The different scoring schemes are depicted in Table 1, exemplified for a CMC item with four subtasks.

Table 1. The different weighting schemes of a CMC item comprising four subtasks

Number of correctly solved subtasks	One point per CMC item	Half point per correct subtask	One point per correct subtask
0	0	0	0
1	0.25	0.5	1
2	0.5	1	2
3	0.75	1.5	3
4	1	2	4

In the first scheme, each CMC item was given a maximum score of one point when all subtasks were solved correctly (*one-point-per-CMC-item* weighting). Hence, in the first model a CMC item was weighted equally to an MC item. In the second weighting scheme, all subtasks of the CMC items were given half the weight of a simple MC item, that is, were scored with half points (*half-point-per-subtask* weighting). In the third scheme, every subtask of a CMC item was awarded with one point and, thus, weighted equally to a simple MC item (*one-point-per-subtask* weighting). Different measures of model fit were considered for evaluating the scoring procedures. The weighted mean square error (WMNSQ, Wright & Masters, 1982) and the respective *t*-value of MC and CMC items were inspected and the information criteria AIC and BIC of the three models were compared.

We then estimated an empirical weight for the two response formats under investigation. To basically reflect the assumption of item homogeneity made with 1PL models, we assumed that all items of the same response formats had the same discrimination. Therefore, we specified 2PL models for polytomous data, also called generalized partial credit models (GPCM; Muraki, 1992) or two-parameter partial credit (2PPC; Yen, 1993) models, in a restricted version. As before, the MC items were scored with one point when answered correctly. The subtasks of each CMC item were aggregated to a polytomous variable, and one point per subtask was awarded. In contrast to the 2PPC with varying item slopes for every item, only two discrimination parameters were estimated: one discrimination parameter for the MC items and one discrimination parameter for the CMC items. For identification reasons, the average of the discrimination parameter for the MC response format was set to one. Consequently, the discrimination parameter of the CMC items in the 2PL model reflected the empirical weight of the CMC response format in comparison to the MC item format.

Results

In the following, we present a) the results of the dimensionality and weighting analyses for scientific literacy in the two different age groups in the NEPS. We then describe the results of the cross-validation analyses b) for ICT literacy in the NEPS, and c) for scientific literacy in PISA.

Scientific Literacy in the NEPS

Dimensionality of the Response Formats. Table 2 depicts the overall fit indices of the uni- and the multidimensional model for the scientific literacy test in G6 and G9. The more parsimonious one-dimensional model suggesting that MC and CMC items form a unidimensional construct was preferred in the G6 scientific literacy test as evident by the lower values of AIC and BIC. In G9 the fit indices exhibited a better fit for the two-dimensional model.

For both age cohorts there were considerable high correlations among the latent variables formed by MC and CMC items (see Table 2). The high correlations in the age cohorts of sixth graders and ninth graders provide strong evidence that the two item formats are measuring the same latent trait.

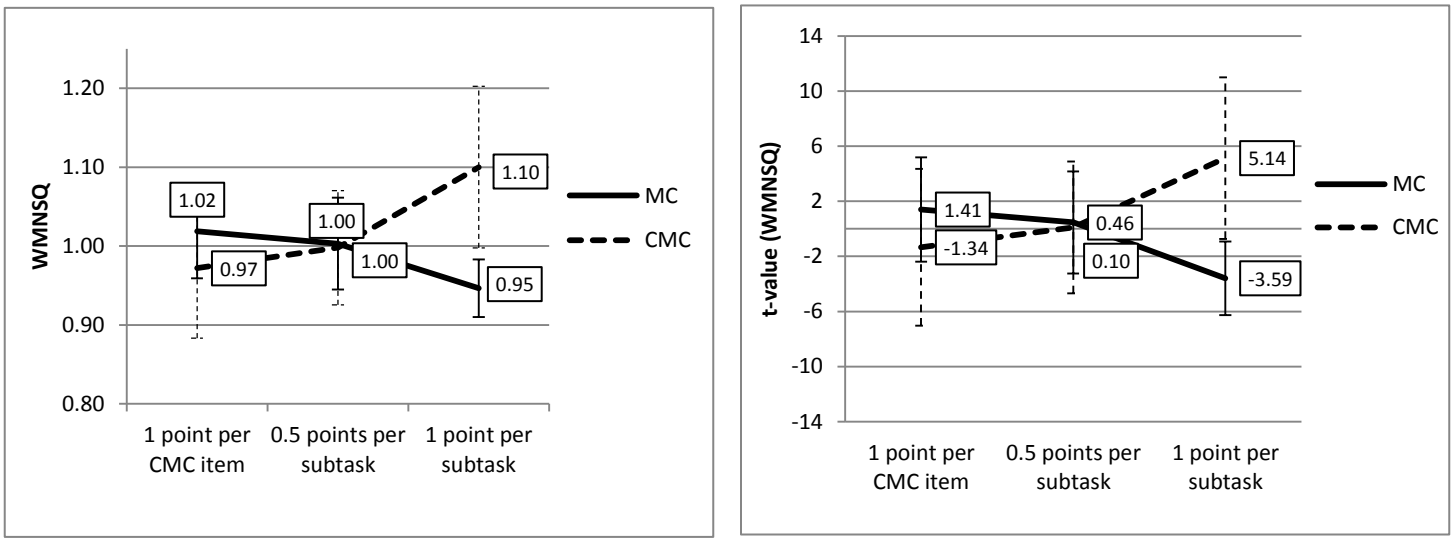
Table 2. Correlation and fit of the uni- and multidimensional models for scientific literacy in the NEPS

Data	Latent correlation	Model	AIC	BIC
G6	0.98	unidimensional	180628.54	180914.15
		two-dimensional	180640.82	180939.41
G9	0.95	unidimensional	580344.03	580752.71
		two-dimensional	580176.06	580599.88

Weighting of the Response Formats. Having endorsed the unidimensionality of the response formats, we investigated which a priori weighting scheme would model the empirical competence data in an appropriate way. The values of the WMNSQ and its *t*-value averaged by the respective response format for the *one-point-per-CMC-item* weighting, the *half-point-per-subtask* weighting, and the *one-point-per-subtask* weighting, are given in Figure 3a and 3b for science in Grade 6 and in Figure 4a and 4b for science in Grade 9. As can be seen in the figures, the average of the WMNSQ and, more evident, the average of the *t*-value for MC and CMC items differed considerably between the weighting schemes. In the G6 scientific literacy test (see Figure 3a and 3b), the one-point-per-CMC-item weighting yielded a slight underfit for the MC items and, conversely, a small overfit for CMC items. In contrast, the one-point-per-subtask weighting resulted in a substantial underfit of CMC items and an overfit of MC items. An almost perfect fit with WMNSQ = 1 for CMC as well as MC items was obtained applying the half-point-per-subtask weighting. Within the response formats, the item fit indices were rather homogeneous for the half-point-per-subtask weighting scheme. For the one-point-per-CMC-subtask weighting scheme, the WMNSQ and the corresponding *t*-values of the CMC items showed greater variance.

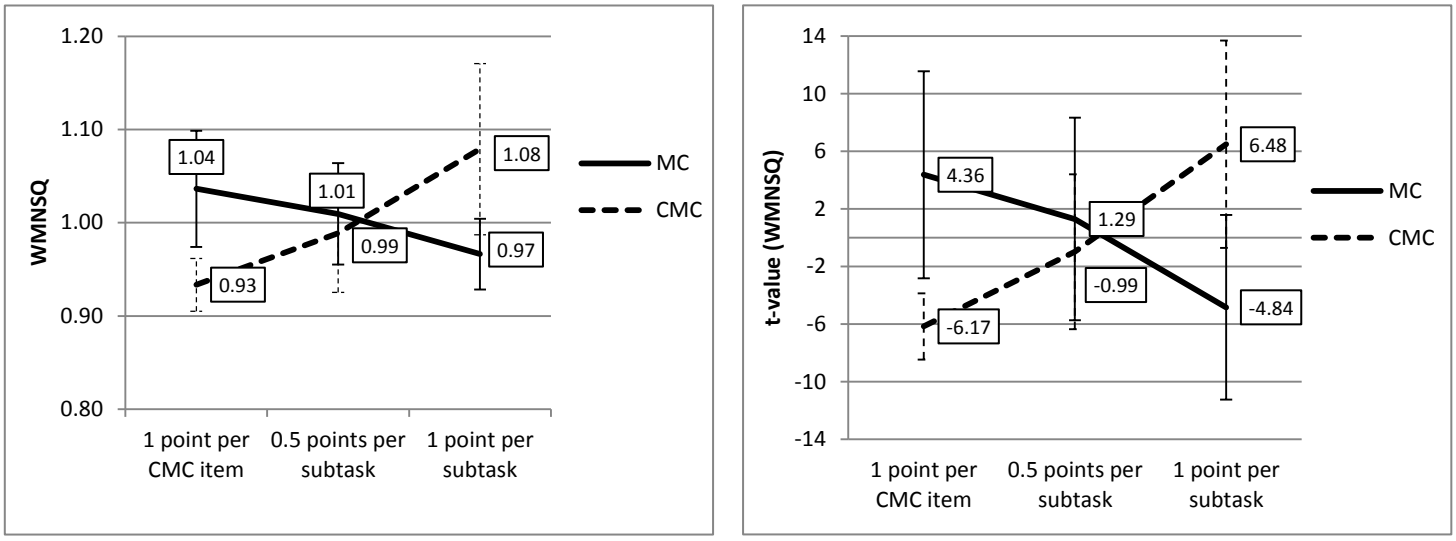
A similar picture of the item fit statistics can be found for the G9 science test (see Figures 4a and 4b). Considerable deviances from an optimal fit for MC and CMC item response formats occurred for the one-point-per-CMC-item weighting scheme and the one-point-per-subtask weighting scheme. The best fit was again achieved when the subtask of the CMC items were scored with half points compared to MC items. Contrary to the G6 science test, the fit indices for the half-point-per-subtask weighting scheme still showed a small underfit of the MC items and a small overfit of the CMC items, indicating that a weighting between half points and one point per

subtask might best approximate the empirical data. Regarding the model fit indices of the three models for G6 and G9 scientific literacy in the NEPS, AIC and BIC values demonstrated a clear preference for the half-point-per-subtask scheme (see Table 3). AIC as well as BIC were smallest when the subtasks of CMC items were awarded half the weight of an MC item.



(a) (b)

Figure 3. Means and standard deviations of (a) the WMNSQ and (b) the *t*-value of the WMNSQ for the three weighting schemes in the G6 science test of the NEPS.



(a) (b)

Figure 4. Means and standard deviations of (a) the WMNSQ and (b) the *t*-value of the WMNSQ for the three weighting schemes in the G9 science test of the NEPS.

Table 3. Fit indices of the models according to the three weighting options for scientific literacy in the NEPS, ICT in the NEPS, and scientific literacy in PISA

Fit criterion	Model	Scientific Literacy G6 NEPS	Scientific literacy G9 NEPS	ICT literacy G9 NEPS	Scientific literacy G9 PISA
AIC	One point per CMC item	181119.36	583667.85	665863.28	7582308.18
	Half point per subtask	180628.54	580344.03	662469.45	7525848.81
	One point per subtask	181962.74	582376.70	665544.09	7536523.32
BIC	One point per CMC item	181404.96	584076.53	666310.56	7582993.78
	Half point per subtask	180914.15	580752.71	662916.72	7526534.40
	One point per subtask	182248.34	582785.38	666109.37	7537208.91

To investigate the empirical weights of the CMC and MC items, restricted 2PPC models were applied to the competence data. The MC items were fixed to have a slope of $a_{MC} = 1$. For the G6 science test, the slope of the CMC items was estimated to be $a_{CMC} = 0.47$. For the G9 science test, the discrimination of CMC items was estimated to be $a_{CMC} = 0.67$. The discrimination for the CMC items in the G9 test above 0.5 corresponded to the item fit indices which had indicated a slight overfit of CMC items for the half-point-per-subtask weighting scheme.

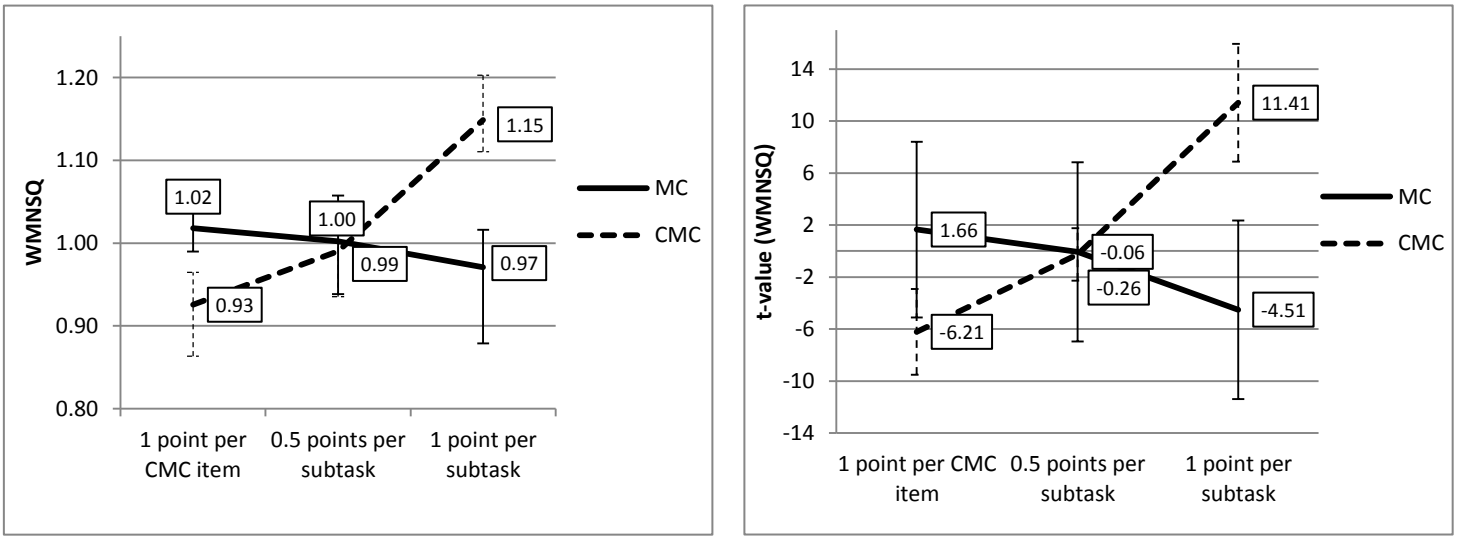
In sum, the results from the scientific literacy tests in different grades in the NEPS provided evidence that the different response formats did not induce sources for multidimensionality, but that they assessed the same underlying competence. Comparing different a priori weighting schemes, weighting subtasks of CMC items with half the weight of an MC item outperformed the other weighting schemes and exhibited a good item and model fit for the tests investigated here. The 2PL analyses revealed that the empirical weights for MC and CMC items were close to the half-point-per-subtask weighting scheme.

Cross-Validation of the Results on an ICT Literacy Test in the NEPS

In order to investigate the generalizability of the results for other competence domains, the same analyses were carried out on NEPS data of an ICT competence test in Grade 9.

Dimensionality. Investigating the dimensionality of the ICT competence test, the descriptive fit criteria indicated a better fit of the two-dimensional model (AIC = 662425.57, BIC = 662888.01) than the unidimensional model (AIC = 662469.45; BIC = 662916.72). We found a latent correlation of $r = 0.96$ between the latent ability based on MC items and the latent ability based on CMC items. The high correlation clearly indicated that the CMC and MC items formed a unidimensional measure.

Weighting. As before, we estimated three 1PL models for ICT literacy based on the different weighting schemes. Figure 5a and 5b depict the average WMNSQ and the *t*-value, separated for MC and CMC items.



(a) (b)

Figure 5. Means and standard deviations of (a) the WMNSQ and (b) the *t*-value of the WMNSQ for the three weighting schemes in the G9 ICT test of the NEPS.

The results indicate that the one-point-per-CMC-item weighting scheme caused a slight underfit of MC items and a substantial overfit of CMC items. Weighting each subtask of CMC item as an MC item enlarged the misfit with a considerable underfit of CMC items and an overfit of MC items. Again, the best fit result was obtained by applying the half-point-per-subtask weighting scheme to the competence data. This was confirmed by the model fit (see Table 3, ICT literacy in the NEPS). AIC and BIC exhibited clear advantages of the half-point-per-subtask weighting rule in contrast to the two other weighting rules.

Having compared the different a priori weighting schemes, we estimated the empirical discrimination indices of the restricted 2PPC model for the two response formats. With the discrimination of the MC items being fixed to $a_{MC} = 1$, the discrimination of CMC items was estimated to be $a_{CMC} = 0.59$. The empirical discrimination, thus, corroborated the half-point-per-subtask scheme and exhibited that the empirical weights were close to the a priori weighting scheme.

Taken together, the results on dimensionality as well as on weighting for ICT competence in the NEPS study replicated the findings for scientific literacy in the NEPS. MC and CMC seemed to measure the same latent ability and the half-point-per-subtask weighting scheme best represented the empirical data.

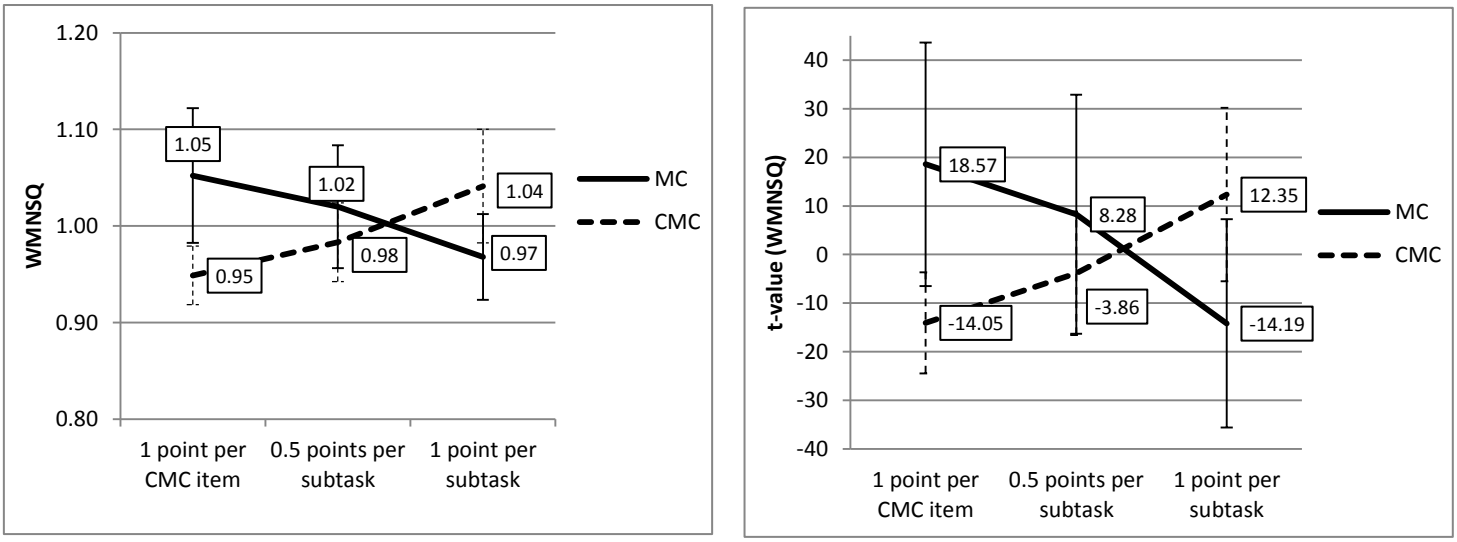
Cross-Validation of the Results on a Scientific Literacy Test from PISA

To augment generalizability of the results across studies, the results were cross-validated on competence data of PISA.

Dimensionality. In the PISA scientific literacy test, the two-dimensional model (AIC = 7520265.54; BIC = 7520972.55) was generally preferred over the unidimensional model (AIC =

7525868.81; BIC = 7526534.40) by the overall fit indices. But again, the latent variables constituted by MC and CMC items were highly correlated (r = .97). The high correlation pointed towards a unidimensional construct measured by the two response formats in the PISA science test.

Weighting. As before, the different weighting schemes for the CMC and MC items were compared in terms of their mean levels of item fit and their model fit. In Figure 6a and 6b the average WMNSQ and corresponding t-values are given for MC and CMC items for each of the three weighting schemes.



(a) (b)

Figure 6. Means and standard deviations of (a) the WMNSQ and (b) the t-value of the WMNSQ for the three weighting schemes in the G9 science test of PISA.

The half-point-per-subtask weighting again resulted in the best fit for both MC and CMC items, although there was still a slight underfit for MC items and, conversely, a small overfit for CMC items. Thus, it is likely that a scoring greater than 0.5 points for the CMC subtasks would best approximate the empirical discrimination of the response formats. The two other scoring

rules yielded a substantial misfit for both MC as well as CMC items. The poorest fit of MC and CMC items occurred for weighting the total CMC items and the MC items equally. AIC as well as BIC (see Table 3, scientific literacy in PISA) had lowest values for the half-point-per-subtask weighting, indicating that the scoring of half points per subtask of a CMC item best captured the empirical competence data.

The estimation of the discrimination of the CMC items in the 2PPC model with the MC items being fixed to $a_{MC} = 1$ was $a_{CMC} = 0.65$. The estimated discrimination of the CMC items suggested an optimal weight of 0.65 for CMC items in a 1PL model. This weight corresponds to the empirical discrimination found for the G9 science test in the NEPS ($a_{CMC} = 0.67$).

In conclusion, the analyses of the PISA competence data also confirmed unidimensionality of the response formats and provided evidence that out of the three a priori weighting schemes the half-point-per-subtask scheme best described the empirical competence data.

Discussion

The current study dealt with the issue of how to appropriately incorporate MC and CMC item response formats in a scaling model. Specifically, we wanted to know whether MC and CMC items that are intended to measure the same construct would empirically form a unidimensional structure. Furthermore, we investigated how well different a priori weighting schemes for the response formats resemble the empirical data.

Examining the dimensionality of the response formats, we found that the results of all competence tests suggested that the two response formats measured the same latent trait. Across age groups, competence domains and studies, the latent correlations of the two dimensions based

on MC and CMC items exceeded $r = .95$, supporting the hypothesis for unidimensionality and justifying a unidimensional scaling of the different item types. We compared these correlations with the latent correlations among the subdimensions of the NEPS scientific literacy and the ICT test that were reported in the working papers on the quality of the test instruments (Schöps & Saß, 2013; Senkbeil & Ihme, 2012). The latent correlations in the G9 science test between the subscales *knowledge about science* and *knowledge of science* were .96, the latent correlations between the subdimensions of ICT ranged from .93 to .96. Hence, the heterogeneity induced by the item response formats was similar or smaller than the multidimensionality emerging from the substantive subdimensions of the domains.

With regard to the cognitive processes associated with the response formats, the results obtained from the present study supported earlier findings on the cognitive facets involved in answering CMC and MC items. The assumption of unidimensionality held across all studies, indicating that MC and CMC items require similar mental processes of recall, recognition, and evaluation. The differences in the MC and CMC response format do not seem to activate different cognitive abilities. We compared the results of the analyses on dimensionality with correlations between MC and CR items from a meta-analysis by Rodriguez (2003) and found that the correlations in the present study were substantially higher. Rodriguez reported corrected true-score correlations of on average $r = 0.85$ across several correlational studies, in which the two item formats were supposed to measure the same trait but the item stems were not equivalent. In the current study, the latent correlations between MC and CMC items ranged between 0.95 and 0.98. These differences in the correlations between MC and CMC and MC and CR items match the distances between the item types in the classification system by Scalise and Gifford (2006). Regarding the degree of constraint in the taxonomy, MC and SR items are considerably more

distant than MC and CMC items. To sum up, the results on dimensionality corroborated the theoretical descriptions of different item types.

The analyses of the a priori weighting schemes consistently demonstrated the advantage of scoring the subtasks of CMC items with half points while allocating one point per correct task for each MC item. The superiority of this weighting rule was persistent across grades (G6 and G9), domains (science and ICT), and studies (NEPS, PISA). The 2PPC models demonstrated empirical discrimination values for the subtasks of CMC items ranging from 0.47 to 0.67. Thus, the estimated discrimination parameters closely resembled the discrimination assumed by the half-point-per-subtask weighting scheme. The reduced empirical discrimination of the CMC subtasks in the present study may arise from the reduced number of response options. Regarding the composition of the two response formats, the number of response options in true-false items constitutes half the number of options of an MC item. Whereas four response options have to be evaluated and compared in MC items, in CMC subtasks there are only two response options requiring these cognitive processes. Thus, CMC subtasks seem to carry about half the information of MC items for the underlying trait.

Allocating one point for CMC items and, hence, equaling them to the MC items or awarding one point per CMC subtask has yielded a considerable over-, or underfit of the MC and CMC items, respectively. Applying the one-point-per-CMC-item weighting rule, we found that substantially more MC items had an unsatisfactory item fit to the model. When applying the one-point-per-subtask weighting rule, the reverse picture occurred. Because items with a poor item fit are often excluded from the final test instrument in the test development process, specific item types might be more likely to be retained when the one-point-per-CMC-item weighting or the one-point-per-subtask weighting is used. Therefore, it seems important to take into account the

impact of weighting different response formats on the item fit when evaluating the items' quality in the process of test construction.

Overall, 2- or 3PL models allow for a more precise modeling of the empirical data, resulting in a better fit of the model to the competence data. However, when a 1PL model type is chosen because of its advantages in allocating theoretical weights for subfacets of the construct, the impact of choosing a weighting scheme for the response formats may be considered. In accordance with the approach in the current study, it may be useful to investigate the relative weight of different response formats at an early stage of test development. On the one hand, a theoretically chosen weighting scheme, for instance, the one-point-per-subtask weighting may be evaluated empirically. When test developers do not have an a priori weighting scheme, they may, on the other hand, estimate empirical weights using restricted 2PL model types. The weights of the response formats can, then, be chosen deliberately for the final scaling model. Considering the determined weights for the response formats, the preferred number of items for the substantive subdimensions of the construct can be chosen to adequately reflect the underlying trait. Subsequently, a sound scaling model may emerge with desirable statistical characteristics and, simultaneously, valuable theoretical features.

Limitations and Directions for Future Research

Altogether, our results seem to generalize to other competence assessments, because relevant factors such as competence domain, grade, or study, have been varied in the present investigation. Moreover, the findings on dimensionality are in line with earlier research pointing to unidimensionality of MC and CMC response formats (Downing et al., 1995; Frisbie & Sweeney, 1982; Hill & Woods, 1974). However, the latent correlations between the response formats in our study were partially higher than the results obtained by Dudley (2006) for a test

assessing second language ability. In conclusion, tests assessing quite different competencies, skills, or abilities might obtain other results for MC and CMC items. Also in competence testings that considerably differ from NEPS and PISA, there may be other response mechanisms and, therefore, other scaling models may be appropriate. In these situations it may be useful to adopt the presented methods and investigate dimensionality and a priori considered item weights of the response formats during test construction and evaluation.

In further studies, it would be valuable to conduct the same analyses on other common item response formats. Innovative item types that were developed only recently (see, e.g., Sireci & Zenisky, 2006) could be implemented to broaden the findings for a wider range of response formats and delineate guidelines for an appropriate implementation in the scaling model. Additionally, further research is needed to study in more detail the psychological processes that are involved in answering the different types of items. The present analyses not only deliver relevant information for scaling models embodying MC and CMC items, but also suggest similar cognitive processes associated with MC and CMC items. By administering tests on cognitive abilities and exploring their relationship to the item formats, more precise conclusions on the mental operations involved could be drawn and cognitive models about the response process of the item formats could be developed.

Acknowledgement

This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort Grade 5, doi:10.5157/NEPS:SC3:2.0.0., Sarting Cohort Grade 9, doi:10.5157/NEPS:SC4:4.0.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

References

Ackerman, T. A., & Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. *Applied Psychological Measurement, 12,* 117-128.

Adams, R., & Wu, M. (2002). *PISA 2000 technical report*. Paris, France: OECD.

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

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

Albanese, M. A., & Sabers, D. L. (1988). Multiple true-false items: A study of interitem correlations, scoring alternatives, and reliability estimation. *Journal of Educational Measurement, 25,* 111-124.

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, Institute of Education Sciences, Department of Education, Office for Educational Research and Improvement.

Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. In N. Brandon-Tuma (Ed.), *Sociological methodology* (pp. 33-80). San Francisco, CA: Jossey-Bass.

Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., et al. (2011). Sampling designs of the National Educational Panel Study: Challenges and solutions. *Zeitschrift für Erziehungswissenschaft, 14,* 51-65.

Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context. *Journal of Marketing Education, 25*, 31-36.

Bennett, R. E., Rock, D. A., Braun, H. I., Frye, D., Spohrer, J. C., & Soloway, E. (1990). The relationship of expert-system scored constrained free-response items to multiple-choice and open-ended items. *Applied Psychological Measurement, 14,* 151-162.

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*, 65-88.

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats – it does make a difference for diagnostic purposes. *Applied Psychological Measurement, 11,* 385-395.

Bleske-Rechek, Zeug, N., & Webb, R. M. (2007). Discrepant performance on multiple-choice and short answer assessments and the relation of performance to general scholastic aptitude. *Assessment and Evaluation in Higher Education, 32*, 89-105.

Blömeke, S., Kaiser, G., & Lehmann, R. (2010). *TEDS-M 2008 – Professionelle Kompetenz und Lerngelegenheiten angehender Primarstufenlehrkräfte im internationalen Vergleich.* Münster, Germany: Waxmann.

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.) (2011). Education as a lifelong process – the German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft, 14*.

Blossfeld, H.-P., von Maurice, J., & Schneider, T. (2011). The National Educational Panel Study: Need, main features, and research potential. *Zeitschrift für Erziehungswissenschaft, 14*, 5-17. doi:10.1007/s11618-011-0178-3

Coderre, S. P., Harasym, P., Mandin, H., & Fick, G. (2004). The impact of two multiple-choice question formats on problem-solving strategies used by novices and experts. *BMC Medical Education, 4*, 23-31.

Downing, S. M., Baranowski, R. A., Grosso, L. J., & Norcini, J. J. (1995). Item type and cognitive ability measured: The validity evidence for multiple true-false items in medical specialty certification. *Applied Measurement in Education, 8*, 187-197.

Dudley, A. (2006). Multiple dichotomous-scored items in second language testing: Investigating the multiple true-false item type under norm-referenced conditions. *Language Testing, 23*, 198-228.

Ebel, R. L. (1970). The case for true-false test items. *School Review, 78,* 373-389.

Ebel. R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Ercikan, K., Schwarz, R., Julian, M., Burket, G., Weber, M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement, 35,* 137-155.

Ferrara, S., Huynh, H., & Michaels, H. (1999). Contextual explanations of local dependence in item clusters in a large-scale hands-on science performance assessment. *Journal of Educational Measurement, 36*, 119-140.

Frisbie, D. A. (1992). The status of multiple true-false testing. *Educational Measurement: Issues and Practices, 5,* 21-26.

Frisbie, D. A., & Druva, C. A. (1986). Estimating the reliability of multiple-choice true-false tests. *Journal of Educational Measurement, 23,* 99-106.

Frisbie, D. A., & Sweeney, D. C. (1982). The relative merits of multiple true–false tests. *Journal of Educational Measurement, 19*, 29-35.

Fuchs, L. S., Fuchs, D., Karns, K., Hamlett, C. L., Dutka, S., & Katzaroff, M. (2000). The importance of providing background information on the structure and scoring of performance assessments. *Applied Measurement in Education, 13,* 1-34.

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

Grosse, M., & Wright, B. D. (1985). Validity and reliability of true-false tests. *Educational and Psychological Measurement, 45,* 1-13.

Guilford, J. P. (1971). *The nature of human intelligence.* London, England: McGraw-Hill.

Haberkorn, K., Pohl, S., Carstensen, C., & Wiegand, E. (in press). Scoring of complex multiple choice items in NEPS competence tests. In H.-P. Blossfeld, J. von Maurice, M. Bayer, & J. Skopek (Eds.), *Methodological issues in longitudinal surveys.* Springer.

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

Haladyna, T. M. (1992). The effectiveness of several multiple-choice formats. *Applied Measurement in Education, 5,* 73-88.

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking.* Boston, MA: Allyn & Bacon.

Haladyna, T. M. (2004). The condition of assessment of student learning in Arizona: 2004. In A. Molnar (Ed.), *The condition of Pre-K-12 education in Arizona: 2004.* Tempe, AZ: Arizona Education Policy Initiative, Education Policy Studies Laboratory, Arizona State University.

Haladyna, T. M., & Downing, S. M. (1989). The validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 1,* 51-78.

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item. *Educational and Psychological Measurement, 53,* 999-1010.

Haladyna, T. M., & Rodriguez, M. C. (2013) *Developing and validating test items*. New York, NY: Routledge.

Hamilton, L. S., Nussbaum, E. M., & Snow, R. S. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education, 10,* 181-200.

Hill, G. C., & Woods, G. T. (1974). Multiple true–false questions. *Education in Chemistry, 11,* 86-87.

Kline, T. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks, CA: Sage.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14*, 117-138.

Manhart, J. J. (1996). *Factor analytic methods for determining whether multiple-choice and constructed-response tests measure the same construct.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, *34*, 207-218.

Martinez, M. E. (1993). Cognitive processing requirements of constructed figural response and multiple-choice items in architecture assessment. *Applied Measurement in Education, 6,* 167-180.

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

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

OECD (2009*). PISA 2006 technical report*. Paris, France: OECD.

OECD (2012). *PISA 2009 technical report*. Paris, France: OECD.

OECD (2013*). PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy.* Paris, France: OECD.

OECD (2014). *PISA 2012 technical report*. Paris, France: OECD.

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

Osterlind, S. J. (1998*). Constructing test items: Multiple-choice, constructed-response, performance, and other formats.* Dordrecht, Netherlands: Kluwer Academic.

Palmer, E. J., & Devitt, P. G. (2007). Assessment of higher order cognitive skills in undergraduate education. Modified essay or multiple-choice questions. *BMC Medical Education, 7,* 49. Retrieved from http://www.biomedcentral.com/1472-6920/7/49/

Penfield, R. D., Myers, N. D, & Wolfe, E. W. (2008). Methods for assessing item, step, and threshold invariance. Polytomous items following the partial credit model. *Educational and Psychological Measurement*, *68,* 717–733.

Pohl, S., & Carstensen, C. H. (2012). *NEPS technical report – Scaling the data of the competence tests.* (NEPS Working Paper No. 14). Bamberg, Germany: University of Bamberg, National Educational Panel Study.

Pohl, S., & Carstensen, C. H. (2013). Scaling the competence tests in the National Educational Panel Study – Many questions, some answers, and further challenges. *Journal of Educational Research Online, 5*, 189–216.

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not reached items in competence tests – Evaluating approaches accounting for missing responses in IRT models. *Educational and Psychological Measurement, 74,* 423-452.

Rodriguez, M. C. (2002). Choosing an item format. In G. Tindal & T.M. Haladyna (Eds.): *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 213-231). Mahwah, NJ: Lawrence Erlbaum.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement, 40,* 163-184.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24,* 3-13.

Rutkowski, L., von Davier, M., & Rutkowski, D. (Eds.) (2014). *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis.* Boca Raton, FL: Chapman & Hall/CRC Press.

Scalise, K., & Gifford, B. (2006). Computer-based assessment in E-learning. A framework for constructing "intermediate constraint" questions and tasks for technology platforms. *Journal of Technology, Learning, and Assessment, 4*. Retrieved [20.10.2014] from http://www.jtla.org

Schöps K., & Saß, S. (2013). *NEPS technical report for science – Scaling results of starting cohort 4 in ninth grade*. (NEPS Working Paper No 23). Bamberg, Germany: University of Bamberg, National Educational Panel Study.

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

Senkbeil, M., & Ihme, J. M. (2012). *NEPS technical report for computer literacy – Scaling results of Starting Cohort 4 in ninth grade* (NEPS Working Paper No. 17). Bamberg, Germany: University of Bamberg, National Educational Panel Study.

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 of Educational Research Online, 5,* 139-161.

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 329-347). Mahwah, NJ: Lawrence Erlbaum Associates.

Snow, R. E. (1993). Construct validity and constructed-response tests. In R. E. Bennett, & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 45-60). Hillsdale, NJ: Lawrence Erlbaum Associates.

Stucky, B. D. (2009). *Item response theory for weighted summed scores* (Master's thesis). Retrieved from https://cdr.lib.unc.edu/indexablecontent?id=uuid:03c49891-0701-47b8-af13-9c1e5b60d52d&ds=DATA_FILE

Thissen, D., Wainer, H., & Wang, X. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement, 31*, 113-123.

Traub, R. E. (1993). On the equivalence of traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 1-27). Hillsdale, NJ: Lawrence Erlbaum Associates.

van den Bergh, H. (1990). On the construct validity of multiple-choice items for reading comprehension. *Applied Psychological Measurement, 14,* 1-12.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Rep. No. RR-05-16). Princeton, NJ: ETS.

Walker, C. M., & Beretvas, S. N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement, 40,* 255-275. doi: 10.1111/j.1745-3984.2003.tb01107.x.

Ward, W. C., Frederiksen, N., & Carlson, S. B. (1980). Construct validity of free-response and machine-scorable forms of a test. *Journal of Educational Measurement, 17,* 11-29.

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. von Maurice (Eds.), *Education as a lifelong process: The German National Educational Panel Study (NEPS) (pp. 67-86).* Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

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

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

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

3.5 Authors' Contributions to the Manuscripts

Haberkorn, K., Lockl, K., Pohl, S., Ebert, S., & Weinert, S. (2014). Metacognitive knowledge in children at early elementary school. *Metacognition and Learning, 9*(3), 239-263, doi:10.1007/s11409-014-9115-1

The new test instrument on metacognitive knowledge was developed under the direction of Kathrin Lockl and Susanne Ebert as part of a subproject (headed by Sabine Weinert) within the interdisciplinary research group BiKS on Educational Processes, Competence Development, And Selection Decisions at Preschool and Elementary School Age. The idea of analyzing the newly developed metacognitive knowledge test with respect to its multidimensionality was developed by Kerstin Haberkorn. The classification systems to be analyzed were chosen and adapted by Kerstin Haberkorn, Kathrin Lockl gave advice. The embedment in the theoretical framework, the specification of the multidimensional models, the specification of the linking model across time, and all data analyses were done by Kerstin Haberkorn. For the analyses, Steffi Pohl gave advice. The idea of investigating the homogeneity of change as further indicator for multidimensionality was introduced by Steffi Pohl. The final specification and application of this model was done by Kerstin Haberkorn. The manuscript was written by Kerstin Haberkorn. The other authors gave advice for revisions.

Pohl, S., Haberkorn, K., & Carstensen, C. (in press). Measuring competencies across the lifespan – challenges of linking test scores. In M. Stemmler, A. von Eye, & W. Wiedermann (Eds.). *Dependent data in social sciences research: Forms, issues, and methods of analysis.* Springer.

Claus Carstensen was responsible for developing the linking design that was used for data collection. The research questions and the analyses schemes were developed by Steffi Pohl. Kerstin Haberkorn and Claus Carstensen contributed to it by discussing them. Kerstin Haberkorn, for example, suggested to also analyze the Grade 7 tests separately. Analyses were done by Steffi Pohl and Kerstin Haberkorn with the help of student assistants. All final analyses with officially released data sets were done by Kerstin Haberkorn. The literature review was done by Steffi Pohl and Kerstin Haberkorn. Specifically, Kerstin Haberkorn contributed to the literature review by reviewing and summarizing common linking practice in (large-scale) studies and results on the coherence of measurement in these studies. The manuscript was mainly written by Steffi Pohl. Kerstin Haberkorn contributed to it by writing a draft for the 'sample and design' as well as the 'measures and procedures' section. She also wrote parts of the theory section and commented and

revised the whole manuscript at all stages of revision. Claus Carstensen commented the paper and added some thoughts in the discussion.

Haberkorn, K., Pohl, S., Carstensen, C., & Wiegand, E. (in press). *Scoring of complex multiple choice items in NEPS competence tests.* In H.-P. Blossfeld, J. von Maurice, M. Bayer, & J. Skopek (Eds.). *Methodological issues in longitudinal surveys.* Springer.

The research questions and the analyses scheme were developed by Steffi Pohl. Kerstin Haberkorn and Elena Wiegand worked on this topic and provided further ideas. A first literature review and first analyses were done by Kerstin Haberkorn and Elena Wiegand under the supervision of Steffi Pohl. The research questions and results for the paper were discussed by Kerstin Haberkorn and Steffi Pohl, Claus Carstensen gave advice. The analyses for the final manuscript, an extended literature review, the embedment in the theoretical framework and the writing of the manuscript were done by Kerstin Haberkorn. Steffi Pohl and Claus Carstensen gave advice for revisions.

Haberkorn, K., Pohl, S., & Carstensen, C. (2015). Incorporating different response formats of competence tests in an IRT model. Manuscript submitted for publication.

The paper builds on the paper by Haberkorn, Pohl, Carstensen, &Wiegand (in press). The idea of analyzing the impact of different a priori weighting schemes on item fit was developed by the three authors. The idea of investigating dimensionality of the response formats was introduced by Kerstin Haberkorn. The literature review and the embedment in the theoretical framework were done by Kerstin Haberkorn. The theory and discussion section with regard to the focus of the manuscript were discussed by the three authors. The idea of investigating the weights of the response formats by additionally applying 2PPC models was introduced by Steffi Pohl and Claus Carstensen. The model was specified and applied by Kerstin Haberkorn. All analyses for the manuscript and the writing of the manuscript were done by Kerstin Haberkorn. Steffi Pohl and Claus Carstensen gave advice.