Psychological Test and Assessment Modeling, Volume 58, 2016 (2), 223-252

Incorporating different response formats of competence tests in an IRT model

Kerstin Haberkorn[1,2]*, Steffi Pohl*[3] *& Claus H. Carstensen*[3]

Abstract

Competence tests within large-scale assessments usually contain various task formats to measure the participants' knowledge. Two response formats that are frequently used are simple multiple choice (MC) items and complex multiple choice (CMC) items. Whereas simple MC items comprise a number of response options with one being correct, CMC items consist of several dichotomous true-false subtasks. 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 items. We cross-validated the findings on another competence domain and on another large-scale assessment. The analyses revealed that the different response formats form a unidimensional measure across contents and studies. Additionally, the a priori weighting scheme of one point for MC items and half points for each subtask of CMC items best modeled the response formats' impact on the competence score and resembled the empirical competence data well.

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

[1] *Correspondence concerning this article should be addressed to:* Dr. Kerstin Haberkorn, Breslaustr. 12, 96052 Bamberg, Germany; email: kerstin.haberkorn@uni-bamberg.de

[2] University of Bamberg, Germany

[3] Free University Berlin, Germany

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.

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. Note that Scalise and Gifford use the term complex multiple choice item (2D.) slightly different to multiple true/false items (2A.) for MC items in which different answers are regrouped into response options. In line with PISA and NEPS, in this paper the term complex multiple choice item is used as synonym to the multiple true/false format (2A) of Scalise and Gifford. 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,

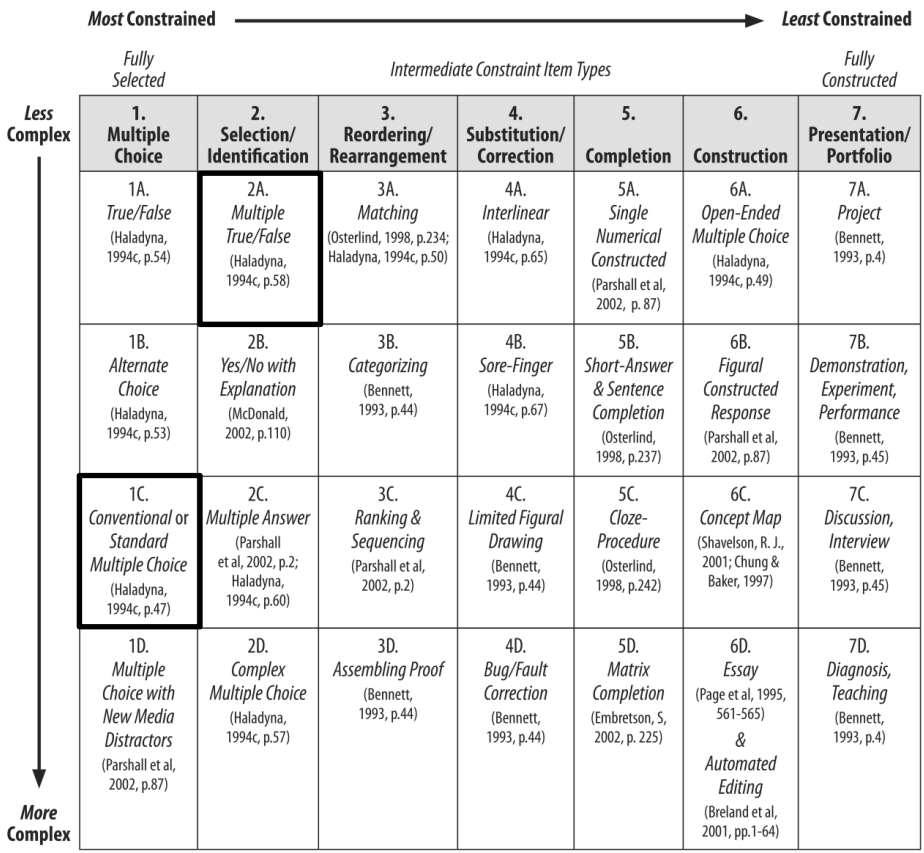


Figure 1:
Classification system for different response formats. The response formats are arranged related to their constraint and their complexity. Adapted from "Computer-based assessment in E-learning. A framework for constructing 'intermediate constraint' questions and tasks for technology platforms" 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 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, 2004a; Hamilton, Nussbaum & Snow, 1997).

The multiple true-false item format (2A.), in this paper termed CMC item format, is placed near the MC format with regard to its constraint. In the CMC item format the choices within the item increase and so the degree of constraint decreases. In contrast to conventional MC items, the subtasks of CMC items demand the subject to mentally generate a counterexample of the response option, because the two response alternatives of the true-false subtasks 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 CMC 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, & Carl-

son, 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 (2003) explored the comparability of SR and CR item formats with variations in item stem and content. For stem-equivalent items, a high average correlation of .95 between the response formats was obtained indicating unidimensionality. Even when the items were not stem-equivalent, but the content to be measured was intended to be the same, latent correlations remained high with an average correlation of .92. Traub (1993) reviewed a number of studies to investigate whether MC and CR items measured the same construct across different domains. He found that the unidimensionality assumption held for MC and CR items in the test instruments assessing reading comprehension and other quantitative domains. In contrast, 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 .89 and .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 theoret-

ically 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.

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

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)

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).

For all analyses with 1PL models, the partial credit model from the family of Rasch models was chosen 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 and partial credit was given for correctly solved subtasks. To avoid possible estimation problems, categories of the CMC items with less than 200 valid responses were subsumed with the adjacent category (Pohl & Carstensen, 2012). This primarily occurred for the lowest two categories of the CMC items. In the G6 science test, the lowest two categories of all CMC items were collapsed except for three CMC items in which three categories were collapsed. In the ICT competence test, the lowest two categories of all CMC items were collapsed. In the G9 science test, the two lowest categories of four CMC items were collapsed. In the PISA test, no categories of CMC items were collapsed.

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 accordance with the scoring in NEPS (Haberkorn, Pohl, Carstensen, & Wiegand, 2015; Pohl & Carstensen, 2012), each category of the polytomous CMC items was scored with half points. 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. 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.

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

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

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

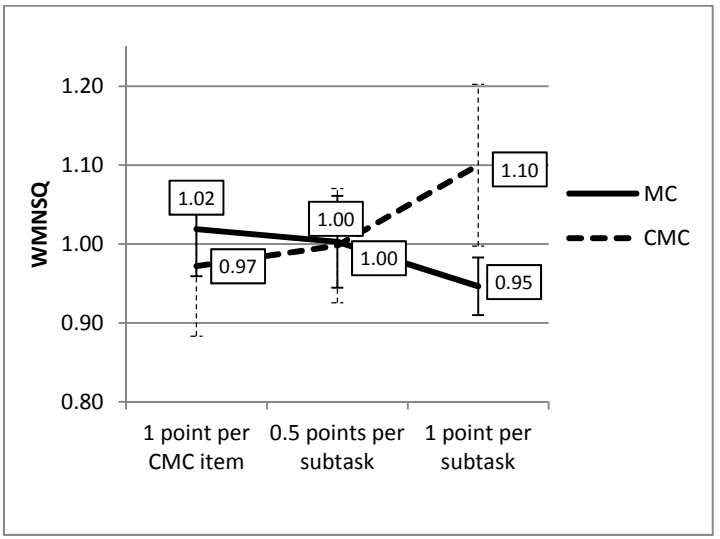
Dimensionality of the response formats. Table 2 depicts the overall fit indices of the uni- and the multidimensional model for the scientific literacy test in G6 and G9. The more parsimonious one-dimensional model suggesting that MC and CMC items form a unidimensional construct was preferred in the G6 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 considerably 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.

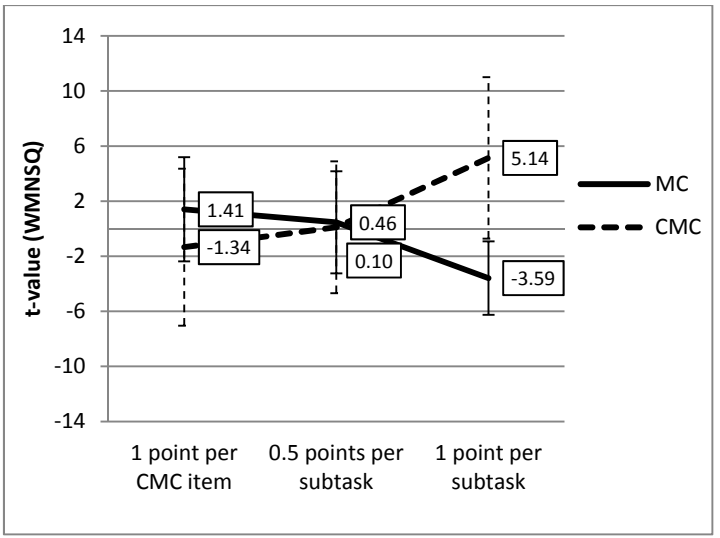
Weighting of the response formats. Having endorsed the unidimensionality of the response formats, we investigated which a priori weighting scheme would model the empirical competence data in an appropriate way. The values of the WMNSQ and its *t*-value averaged by the respective response format for the *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.

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

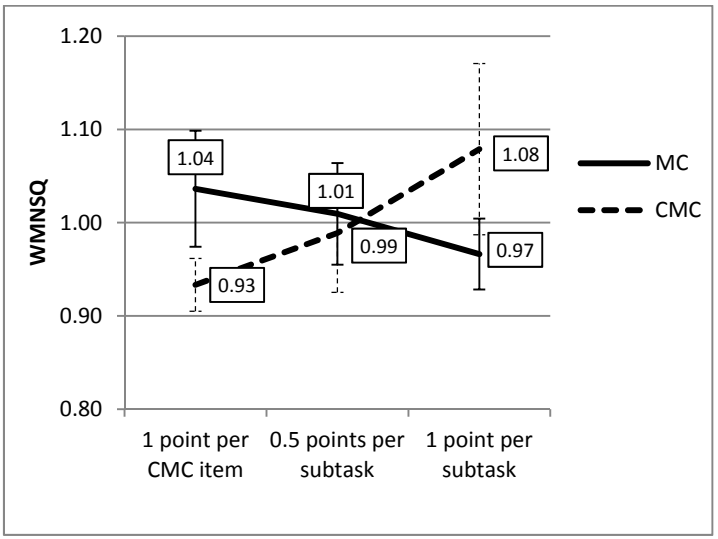


(a)

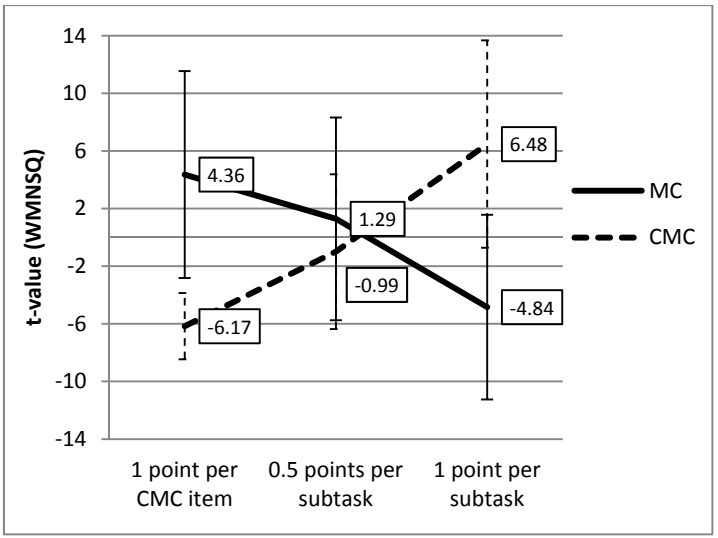


(b)

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

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.

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 $a_{CMC} =$

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

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.

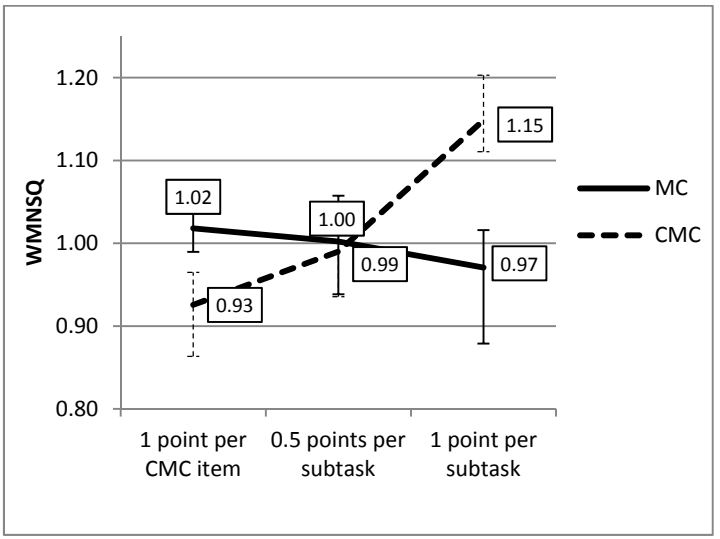
Dimensionality. Investigating the dimensionality of the ICT competence test, the descriptive fit criteria indicated a better fit of the two-dimensional model (AIC = 662425.57, BIC = 662888.01) than the unidimensional model (AIC = 662469.45; BIC = 662916.72). We found a latent correlation of $r = .96$ 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.

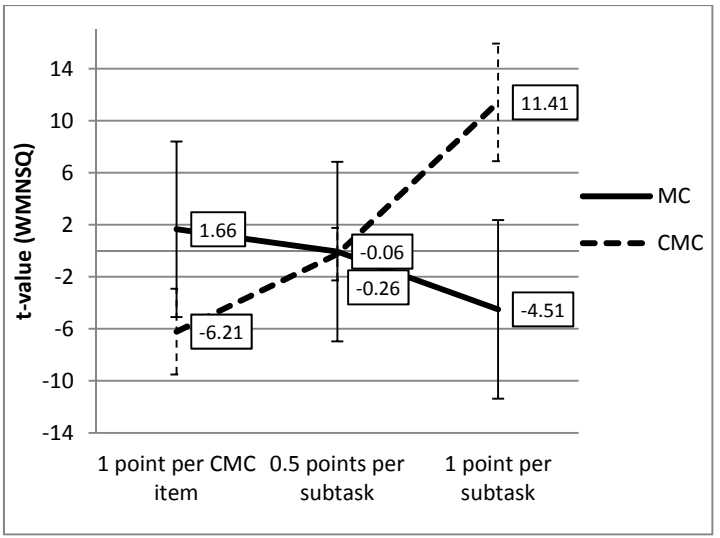
The results indicate that the one-point-per-CMC-item weighting scheme caused a slight underfit of MC items and a substantial overfit of CMC items. Weighting each subtask of CMC item as an MC item enlarged the misfit with a considerable underfit of CMC items and an overfit of MC items. Again, the best 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.



(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.

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.

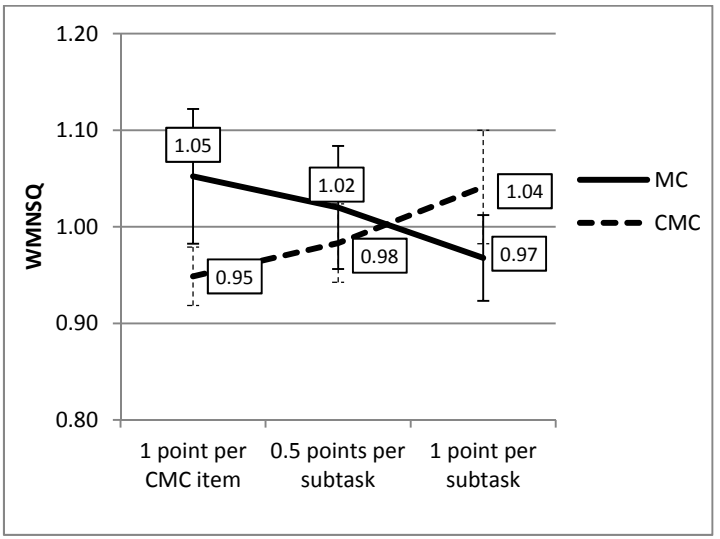
Dimensionality. In the PISA scientific literacy test, the two-dimensional model (AIC = 7520265.54; BIC = 7520972.55) was generally preferred over the unidimensional model (AIC = 7525868.81; BIC = 7526534.40) by the overall fit indices. But again, the latent variables constituted by MC and CMC items were 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.

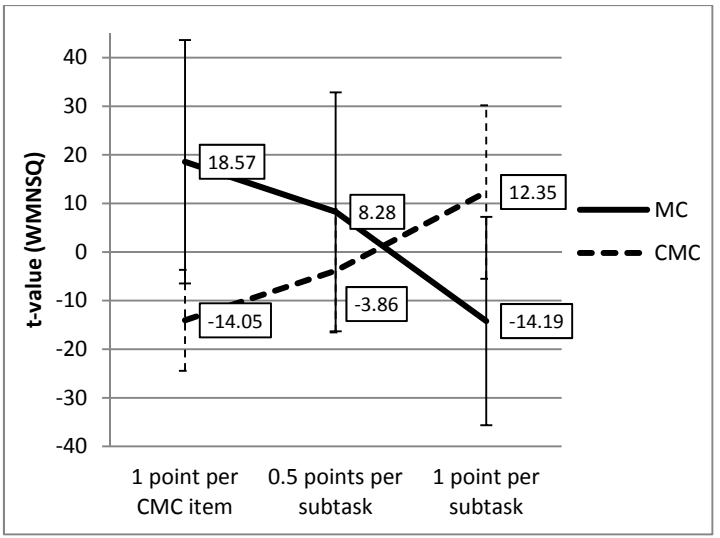
The half-point-per-subtask weighting again resulted in the best fit for both MC and CMC items, although there was still a slight underfit for MC items and, conversely, a small overfit for CMC items. Thus, it is likely that a scoring greater than 0.5 points for the CMC subtasks would best approximate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



(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

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 higher. Rodriguez reported corrected true-score correlations of on average $r = .92$ 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 .95 and .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 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 the true-false subtasks of the CMC 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. Moreover, the incidence of guessing within the CMC subtasks (Haladyna, 2004b) is higher, as they only consist of two response options. This might additionally reduce the information that can be gained from them for the overall competence score. On the other hand, Haladyna, Downing, and Rodriguez (2002) stated that with an increasing number of subtasks per CMC item, the influence of guessing can be reduced. In conclusion, the CMC subtasks seemed 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 our study, some categories of the polytomous CMC items were collapsed because they had less than 200 valid responses. Overall, collapsing of categories may lead to a loss of information and may bias parameter estimates (e. g. Ben-Simon et al., 1997). However, we did not assume a considerable bias in our study, since mainly only two categories within the CMC items were collapsed. Moreover, collapsing only affected a small number of subjects (less than 200) per CMC item. In further analyses, we found no systematic relationship between the collapsing of categories and empirical weights of the CMC items. Yet, in the development of scaling models it seems relevant to analyze effects of subsuming categories on other parameter estimates in order to not confound the results.

In some of the competence tests under investigation, the CMC items differed in their number of subtasks ranging from three to seven subtasks. Thus, in subsequent analyses we explored whether the number of CMC subtasks and the empirical weights of the CMC items were related. However, no consistent relationship was found across studies. In order to confirm whether these findings may be generalized, one future research task should be to analyze this relationship using CMC items which differ more in their number of subtasks.

In further studies, it would also 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 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.

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. (2016). 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* (pp. 523-540). Wiesbaden: 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. (2004a). 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. (2004b). *Developing and validating multiple-choice test items*. Mahwah, NJ: Lawrence Erlbaum Associates.

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., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15,* 309-334.

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. (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.