Scoring of Complex Multiple Choice Items in NEPS Competence Tests

Kerstin Haberkorn, Steffi Pohl, Claus Carstensen and Elena Wiegand

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.

Hans-Peter Blossfeld et al. (eds.), *Methodological Issues of Longitudinal Surveys*,
DOI 10.1007/978-3-658-11994-2_29

1 Item Formats and Scaling Model of the NEPS Competence Tests

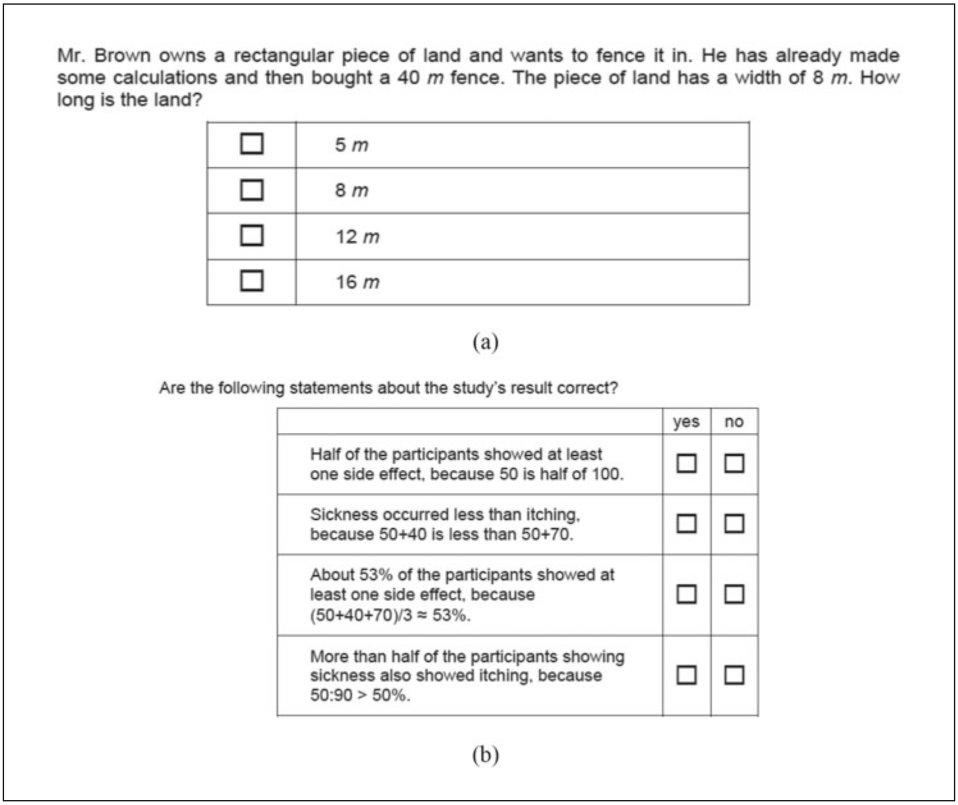
In the process of test development, the choice of the items' format plays a crucial role for different aspects of validity (Rodriguez, 2002). So far, comprehensive item writing rules and guidelines have been published (Downing & Haladyna, 2006; Haladyna & Rodriguez, 2013; Osterlind, 1998), and a variety 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) items[1] (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.

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

[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.

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



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)

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.[2] 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 vari-

[2] 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.

able 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 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 1, the item difficulty and discrimination for the All-or-Nothing scoring and the NC scoring in the Science and ICT domains are depicted.

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 oc-

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 curred. 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). 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.

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 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.[3] 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 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

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

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) 0.5 points per correct subtask**	**(c) 1 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

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.

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.

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 maxi-

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

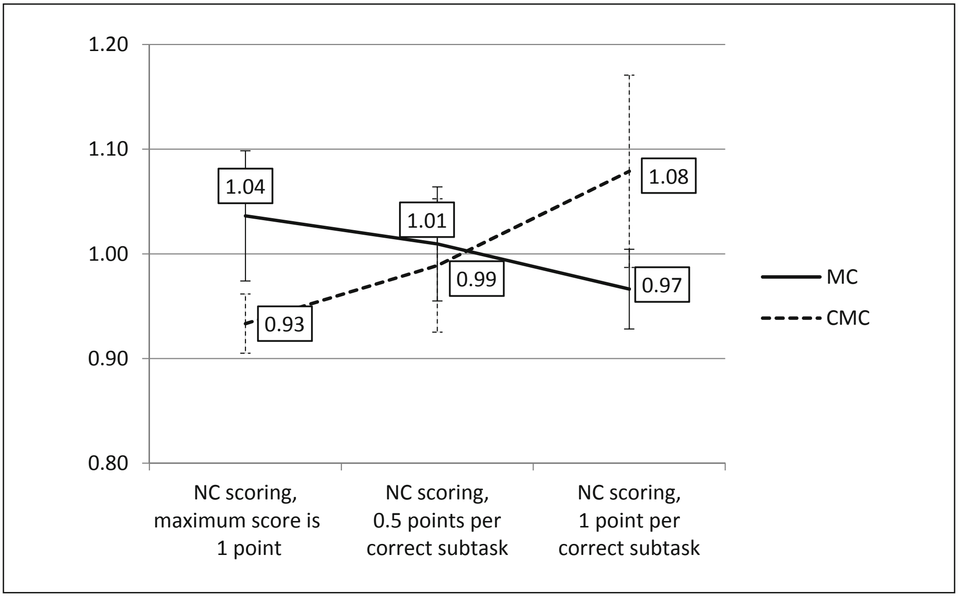
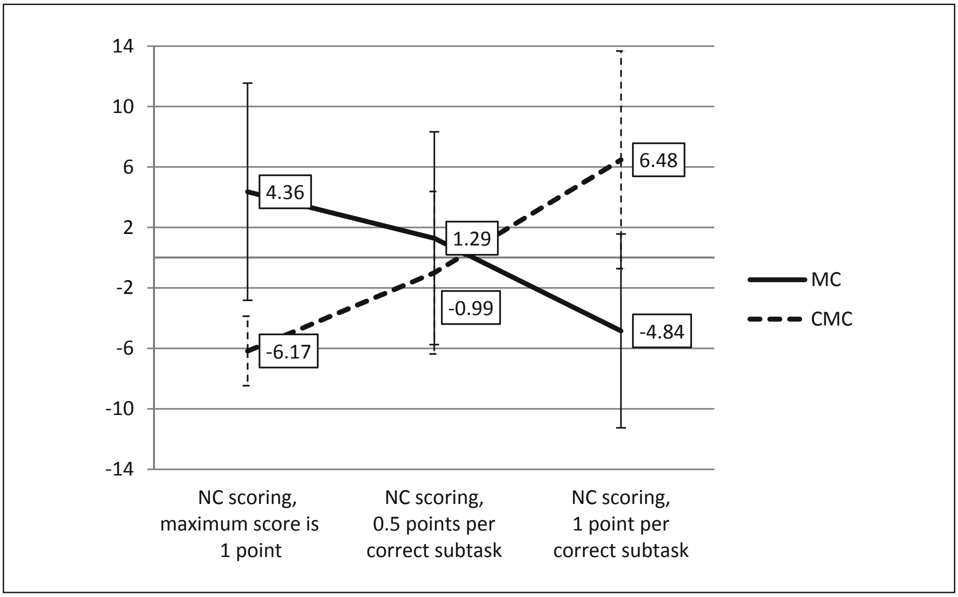


Figure 3 Means and standard deviations of the *t*-value of the WMNSQ for different item weightings in the domain of Science (Haberkorn et al., 2015)



mum 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).

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 com-

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, 0.5 points per correct subtask	NC scoring, 1 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.

petence 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.

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: 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., … 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*(2), 110–138.

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

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*(1), 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*(2), 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*(2), 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*(2), 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*(3), 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*(2), 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*(4), 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), 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*(3), 187–213.

About the authors

K. Haberkorn
University of Bamberg, Bamberg.
e-mail: kerstin.haberkorn@uni-bamberg.de

C. Carstensen
Leibniz Institute for Educational Trajectories (LIfBi), Bamberg.
Chair of Department for Psychology and Methods of Educational Research,
University of Bamberg, Bamberg.

S. Pohl
Chair of Methods and Evaluation/Quality Assurance,
Free University Berlin, Berlin.

E. Wiegand
University of Mannheim, Mannheim.