

## How to measure critical health competences: development and validation of the Critical Health Competence Test (CHC Test)

Anke Steckelberg · Christian Hülfenhaus · Jürgen Kasper · Jürgen Rost ·  
Ingrid Mühlhauser

Received: 21 May 2007 / Accepted: 12 September 2007  
© Springer Science+Business Media B.V. 2007

**Abstract** Consumers' autonomy regarding health increasingly requires competences to critically appraise health information. Critical health literacy refers to the concept of evidence-based medicine. Instruments to measure these competences in curriculum evaluation and surveys are lacking. We aimed to develop and validate an instrument to measure critical health competences (CHC test). Development and testing of the questionnaire covered three phases: (1) test construction (*and feasibility*), (2) first field test of scalability and items revision (3) second field test to validate the instrument. Model fit analyses were performed for both field tests for Rasch-, Mixed Rasch- and Hybrid model. Participants were secondary school and university students with and without prior training in evidence-based medicine (1. field test  $n = 322$ ; with training  $n = 87$ ; 2. field test  $n = 107$ ; with training  $n = 13$ ). The second field test resulted in Rasch scalability of all items in one person class. Mean values ( $\pm$ SD) of person parameters were 716.14 ( $\pm$ 53.74) for trained students and 470.11 ( $\pm$ 59.63) for untrained students. Reliability of the instrument was 0.91 (WINMIRA ANOVA). In conclusion the CHC instrument is a feasible, reliable and valid instrument to measure critical health literacy. The generalizability of the instrument is to be explored in ongoing studies in different educational settings.

**Keywords** Competence test · Evidence-based medicine · Health education ·  
Health literacy

---

A. Steckelberg (✉) · I. Mühlhauser  
Unit of Health Sciences and Education, University of Hamburg,  
Martin-Luther-King Platz 6, 20146 Hamburg, Germany  
e-mail: asteckelberg@uni-hamburg.de

C. Hülfenhaus · J. Kasper  
Unit of Health Sciences and Education, University of Hamburg,  
Martin-Luther-King Platz 6, 20146 Hamburg, Germany

J. Kasper  
Institute of Neuroimmunology and Clinical MS Research (INiMS), University of Hamburg,  
Hamburg, Germany

J. Rost  
Leibniz-Institute for Science Education, Christian-Albrechts University of Kiel, Kiel, Germany

## CHC test

The enhancement of health literacy, internationally, is considered a public goal and a challenge for health education in the 21st century (Nutbeam 2000). Patients and consumers are increasingly taking responsibility for diagnostic and therapeutic decisions without being prepared for a critical assessment of the various procedures and products (Taking health literacy seriously 2005).

Health literacy was first defined in the USA in 1974 as, “The degree, to which individuals have the capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions.” (Simonds 1977).

The extended definition underlying the World Health Organizations’ (WHO) healthy schools initiative defines health literacy as follows: “Health literacy represents the cognitive and social skills, which determine the motivation and ability of individuals to gain access to understand and use information in ways which promote and maintain good health.” (European Network 2004; Lynagh et al. 2002). However, the WHO approach remains paternalistic. The initiatives aim at compliance towards pre-defined objectives instead of informed choices.

Teaching competences that enhance consumers’ autonomy regarding health requires combining the two concepts of health literacy and evidence-based medicine/evidence-based health care (EBM). Competences originating from this new concept are referred to as critical health literacy and are subject of the test development presented in this contribution.

The objective of literacy assessment in educational practice addresses the comprehension of concepts and the ability to deal with different situations (Deutsches PISA Konsortium 2001; Max Planck Institute 2000). Competence trainings that are based on the concept of EBM and that aim to promote critical health literacy are primarily provided for physicians. There are only few isolated projects in the USA, Great Britain and Germany that train patients and patients’ advocacies with regard to critical health literacy (Dickersin et al. 2001; Milne and Oliver 1996; Rosenfeld et al. 2002; Berger et al. 2007).

Most available assessment instruments target EBM training programs for physicians and evaluate knowledge, skills, attitudes and behaviors (Shaneyfelt et al. 2006). Therefore, these instruments are not appropriate to assess critical health competences of health care system users.

However, as a novelty, in a recent pilot study we developed and piloted a curriculum for secondary school students, which aims at the enhancement of critical health literacy.

The purpose of this study was to develop and validate an instrument for measuring critical health competences. The instrument should consider two different conditions. On the one hand it has to be specific enough to be applied in an evaluation study of a particular curriculum. On the other hand it is supposed to measure critical health competences in surveys.

## Development process of the CHC test including methods and results

We developed a questionnaire, which was to be an adequate method for measuring large samples of secondary school students and health care users in a manageable amount of time. Since the instrument has to be applicable to curriculum evaluation as well as to surveys of competence levels, a measure is needed, being able to assess varying levels of competence with comparable reliability. This condition holds for scales constructed in

terms of the Rasch model (Rost 2004). The study aimed at developing a Rasch scaleable competence measure, because item difficulties and person abilities are conceived as independent. Furthermore, as a property of the Rasch model, the person parameters can be assessed with the entire test instrument or shortened screening versions.

The development and pre-testing of this questionnaire covered three pre-defined phases of collecting empirical data:

Phase 1: A first version of the questionnaire was constructed and pre-tested by collecting qualitative data from 8 students. In a face-to-face setting; the students were observed and interviewed with a focus on their understanding of the questions and the response formats.

Phase 2: After revising the test according to the results of the first phase, a quantitative field test was performed. This first field test aimed at getting information about the fit of the Rasch model to this kind of competency data and the appropriateness of the facet design of the test. Additionally, the response format, distractors, item difficulties and discriminations were analyzed, and the test instrument has been revised accordingly.

Phase 3: A second field test was performed in order to control the improvement of the test instrument and the fit of the Rasch model. The development process was completed by the time the Rasch model was the best fitting model.

## Phase 1: test construction

### *Facet structure*

A 4 by 4 facets design was employed to define the item domain and to guide our item construction. The first facet consisted of four different content areas of health care that were known as being topics of high interest in the target group (Steckelberg et al. 2006) 1. Echinacea and common cold; 2. Magnetic resonance imaging in knee injuries; 3. Treatment of acne; and 4. Breast cancer screening. The topics were transformed into scenarios that were described at the beginning of each unit and built the context of item formulation.

The second facet was built up by four subareas of competence, representing the underpinning theoretical structure of the critical health literacy construct. These categories are: A. Understanding medical concepts; B. Skills of searching literature; C. Basic statistics; and D. Design of experiments and sampling (Table 1). This second facet built the framework for constructing similar or even parallel items for each of the four scenarios.

The item domain is exhaustively defined by these two facets: The items should cover major aspects of all four categories of both facets, i.e. 16 categories of items were to be filled with a minimum of two and a maximum of 7 items. Wording of the items had to

**Table 1** Facet structure

Categories	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Sum
A: Medical concepts	2	5	2	6	15
B: Literature	7	6	5	4	22
C: Statistics	2	5	4	7	18
D: Design of experiments and sampling	5	4	4	4	17
Number of items	16	20	15	21	72

consider two conditions: Reference of items to a particular scenario and exchangeability between the scenarios.

A first pool of test items was developed by AS, CH, and AF independently of each other, and covered 250 items. The project team (AS, CH, and AF) evaluated these items with respect to completeness, duplicates, readability, and comprehensibility. Selecting the best items according to these criteria resulted in a final pool of 72 items. High drop out rate was due to duplicates.

### *Construction of test*

The items were grouped according to the 4 scenarios described above. According to the second facet (type of knowledge) an unsorted order was applied within the units. The multiple choice and open-ended questions were evenly distributed.

The resulting four test sheets were reviewed again (AS, CH, JK, JR) regarding the reference of each single item to the scenario, response format (open or multiple choice), clearness of distractors and expected response time. The application of all scenarios should not exceed 90 min.

### *Coding instructions*

In order to ensure objective analysis of the tests, coding rules for free item responses and scoring rules for multiple choice items were defined for every item, which is especially important for open formats. Coders were trained to stick to the coding instructions and the agreement of different raters coding the same responses was calculated. In case of too low agreement, either the coding instructions were revised or the raters were trained again until interrater-reliability reached a level of 95–100%.

### *Pre-tests*

Eight students of the target population were selected for testing feasibility. Six of them answered each item individually, using a think-aloud protocol to articulate their understanding of the item, their choice of answer, and why each of the other choices was eliminated as a possible correct answer. The think-aloud protocol helped to identify and remove language problems and conceptual constraints. Revisions were made after each test application. In addition, we finally had 2 students to work on the items without further instructions to get information on necessary timing.

Phase 2: first field test

### *Test administration*

Students were asked to respond to the four test sheets within 90 min. To prevent sequential effects, the order of the four test sheets varied randomly. Calculators were permitted. The participants were assured that data survey and processing would be performed anonymously.

### Sample and setting

For a first field test of the new instrument, comprising 72 items, 300 students had to be surveyed. Starting with a somewhat higher sample size because of the expected shrinking rate, the final sample comprised 322 trained and non-trained secondary school students (grade 10 and 11) and university students. The latter subsample was recruited from different courses on evidence-based medicine of the study course Health Sciences at the University of Hamburg. We documented age, sex, mother-tongue, and further languages, type of school and participation in competence training.

### Results

Descriptive data of the sample are shown in Table 2. Subsamples of secondary school students showed little differences in sex. The smaller proportion of girls in the trained subsample is due to one of the pilot schools, which focuses on science education and therefore attracts more boys. The difference in first language distribution is due to our selection criteria for the pilot courses. We intended diverseness of participating classes regarding mixture of students, pre-conditions of learning and proficiency level. The high proportion of female university students corresponds to the distribution in the educational program of health sciences.

Item analysis procedures started with distractor and missing value analyses of the multiple choice items using SPSS 13.0. Correlation patterns of single attractors and distractors were consistent with the theory. Further model fit analyses were calculated using WINMIRA 2001, which is a software for analyses with a variety of discrete mixture distribution models for dichotomous and polytomous categorical data and can be used for the Rasch Model, the Latent Class Analysis, the Mixed Rasch Model and the Hybrid Model (Davies 2001).

### Model fit testing

To achieve Rasch scalability of the test and the students was one of the primary goals in this study. So our first question that had to be answered by the data referred to the property of homogeneity of items and persons. We investigated, whether Rasch scalability of the test was given in the entire target group of persons or only for an (unknown) subpopulation. The question of person homogeneity can be investigated by mixture distribution models.

**Table 2** Sample characteristics (First field test)

	Students (secondary school)		Students (university) ( <i>n</i> = 67)
	With training ( <i>n</i> = 37)	Without training ( <i>n</i> = 218)	
Age in years <sup>a</sup> mean values (SEM)	17.61 (0.13)	17.37 (0.07)	27.16 (0.62)
Sex (female)	20 (56%)	138 (64%)	46 (75%)
First language German	28 (78%)	182 (84%)	59 (77%)

Note. data are shown as absolute numbers (valid percentages)

<sup>a</sup> Missing values

We applied the Mixed-Rasch model, which extends the Rasch model to a discrete mixture model. Applying this model, an inhomogeneous sample of persons can be divided (unmixed) into Rasch-homogenous subsamples, called latent classes. In our analysis, only two-classes mixed Rasch models have been calculated in order to compare the fit of these models with the Rasch model.

As a second alternative to the Rasch model we applied a Hybrid Model, which assumes that only one class of persons can be measured by the Rasch model, whereas the other class(es) is/are defined by a set of constant response probabilities. Hybrid models can be described as a combination of a Rasch Model and Latent Class Analysis (Yamamoto 1989). Applied to questionnaire data, the Hybrid model identifies for each observed response pattern whether it belongs to a latent subpopulation where the Rasch model holds or to a subpopulation, where scaling is not possible. Compared to both, the mixed Rasch and the Hybrid model, the (simple) Rasch model is the restricted case of only one latent class, in which all persons are homogeneous and, hence, Rasch scalable.

Model fit analyses were separately conducted for the four scenarios. Beside the Rasch model, we also tested the Mixed-Rasch model and the Hybrid model (2-classes in each model). Model fit was evaluated by means of information criteria. The Bayes Information criterion (BIC) is based on the log-likelihood of the data, given a particular model, and the number of estimated model parameters. The number of parameters operates like a penalty term, so that more parsimonious models are preferred. The BIC-values of each test (scenario) were compared between the competing models, to ascertain the best fit.

Table 3 shows the BIC-values of all models and the percentage of participants that belong to the Rasch-class. The model with the smallest BIC had to be chosen. Table 3 shows that the Rasch model was not the best fitting model with none of the 4 scenarios. It turns out that, according to the BIC, the Hybrid model is the best fitting model among the three compared models.

This result raises the question, whether more or less the same students belong to the class of Rasch-scalable persons in all 4 scenarios. This question may be answered by means of a 2-by-2-by-2-by-2 cross table or by applying an ordinary Latent Class Analysis to the dichotomous class membership (1 = person is Rasch scalable, 0 = is not) of all students.

The resulting two class model supports the hypothesis of consistent class membership over all four scenarios. Students affiliated with the first class are classified in the non-scalable class in scenario 1–4 with probabilities ranging from 80–91%. The second class comprises students with high probabilities to be classified in the scalable class in all scenarios with probabilities ranging from 81 to 96%. High consistency of class affiliations was also shown by high mean probabilities of expected (meta-) class memberships of 98% in both classes.

**Table 3** First field test: BIC scores and class distribution

Analyses	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Overall
Rasch (1 class)	4959.56	6729.50	4178.55	7411.00	22779.37
MiRa (2 classes)	4915.55	6768.78	4201.51	7459.15	22725.65
Class 1	65%	75%	70%	74%	
Hybrid model	4875.04	6671.85	4147.00	7378.75	22486.80
LCA Class	62%	73%	69%	64%	

Note. MiRa = Mixed Rasch Model; LCA = Latent Class Analysis

Reliability (WINMIRA ANOVA) across all items for the Rasch model was 0.88 and for single scenarios 0.74 (scenario 1); 0.68 (scenario 2); 0.69 (scenario 3) and 0.74 (scenario 4).

In order to show effects of the curriculum on students who had taken part in the competence training at school we reported person parameters for the subsamples of secondary school students. In addition person parameters of university students are reported (Table 4). The effect sizes of the subsamples were 1.32 (95% CI 0.95–1.68) (Cohen's *d*) for secondary school students, indicating a large effect. For the university students Cohen's *d* was 0.64 (95% CI 0.08–1.2), indicating a medium effect. Cohen's *d* for untrained secondary school students and trained university students was 3.05 (95% CI 2.65–3.45). Although the effect of school type seems to be higher than the effect of training, the interaction of both factors is not statistically significant. This result confirms our hypotheses on the effectiveness of the training for any age group.

Since Rasch scalability was the aim of the development process, we based the revision process of the items on the Rasch results of this first field test, which showed that scalability was not yet achieved.

### Item revision

The item revision aimed at reaching Rasch scalability. Item revision was performed according to the following criteria derived from the Rasch model analysis:

1. All item-Zq values were checked. High positive values indicate lower discrimination than expected (Rost and von Davier 1994). All significant *p*-values ( $p < 0.05$ ) indicating misfitting items were considered as revisable.
2. Item scores were checked for response probability, considering items <15% response probability as revisable, since overall response probability should exceed 40% in order to eliminate bottom effects.
3. Distractors and attractors were reviewed for frequencies of ticks. Attractors that were rarely marked were checked for plausibility as well as distractors that were marked very often, leading to false answers.

Numbers of items identified for revision are shown in Table 5.

Item revision was performed by changing the format, replacing distractors or adding further information or modifying coding instructions. Two items were removed for very low item scores with small standard deviations indicating low discriminatory power. All in all we revised 51 items and added 6 new items. Three items that were identified as revisable were left unmodified, since they were regarded as important for testing high performing students. Coding instructions were adjusted correspondingly.

**Table 4** Person parameter of students

	<i>n</i>	Person parameter <sup>a</sup>	SD
Untrained students (secondary school)	218	451.75	65.29
Trained students (secondary school)	37	530.98	54.69
Untrained students (university)	17	606.77	73.79
Trained students (university)	50	651.16	65.57

<sup>a</sup> Values are means

**Table 5** Items identified for revision

Scenario	Criteria for item revision				Revised	Results	
	Zq value	Item score	Distractors	Elimination		Still revisable	
						Zq value	Item score
Scenario 1	4	5	5	0	5	6	3
Scenario 2	6	5	7	1	10	2	3
Scenario 3	4	4	2	0	5	1	3
Scenario 4	5	7	8	0	7	1	2

### Phase 3: second field test

#### *Sample and setting*

The sample ( $n = 107$ ) comprised four grade 11 secondary school classes and students of the University of Hamburg, Unit of Health Sciences and Education. *Test administration.* Again students were asked to respond to all four test sheets within 90 min. All procedures were identical to the first field test.

#### *Results*

Descriptive data of the sample are shown in Table 6. The difference in age is due to two older students (41 and 42 years). The lack of male students is not representative. Compared to the sample of the first field test, untrained secondary school students are comparable in both tests.

To assess objectivity of the transfer of data into the database, including multiple choice and open-ended questions, the interrater-reliability was calculated between the codes for each single item distractor ( $n = 313$ ) of two trained coders rating 30 datasets independently. Cohen's Kappa was excellent for 277 ratings ( $\kappa = 0.9\text{--}1.0$ ), moderate or good for 31 ratings ( $\kappa = 0.7\text{--}0.89$ ) and poor for 5 ratings ( $\kappa = <0.7$ ), resulting in additional information in the manual of the test.

#### *Model fit testing*

The second field test again investigated Rasch scalability of the revised test. We expected to get Rasch scalability for all items and persons (i.e. only one latent class).

**Table 6** Sample characteristics (Second field test)

	Students (secondary school) ( $n = 94$ )	Students (university) ( $n = 13$ )
Age in years <sup>a</sup> (SEM)	16.5 ( $\pm 0.1$ )	30.6 ( $\pm 1.4$ )
Sex female <sup>b</sup>	59 (63%)	13 (100%)
First language German <sup>b</sup>	84 (89%)	13 (100%)

<sup>a</sup> Values are means (SEM)

<sup>b</sup> Values are absolute numbers (valid percentages)

Table 7 shows the “Bayes Information Criteria” (BIC) for the Rasch, the mixed Rasch and the hybrid model. The one class Rasch model turned out to be the “best fitting model”, for its BIC scores were the lowest for all scenarios and also for the whole sample (Table 7).

Overall response probability reached 0.34 as compared to 0.27 in the first field test. The mean response probabilities of the single scenarios changed in the following way: scenario 1: 0.27 → 0.38; scenario 2: 0.27 → 0.34; scenario 3: 0.27 → 0.30 and scenario 4: 0.28 → 0.34, so that the bottom effect was reduced. Some items are still very difficult, in particular in samples that were not trained or for samples at the secondary school level.

This gives reason for investigating, if there are systematic differences in the item difficulties between the scenarios or the item types (second facet). Item parameter estimates for all types of knowledge and cognitive operations are depicted in Fig. 1. This figure shows that the difficulty parameters in both field tests were lowest for the first type ‘understanding medical concepts’ and highest for the third item type ‘basic concepts of statistic’. In particular, items of the first type were improved, i.e. became easier by the item revision. Figure 1 shows, that all items in this category were easier in the second field test, except for the third and the last two items.

The differences of item difficulties between the scenarios are not represented by the item parameters, because the Rasch analyses were performed separately for the scenarios and the item parameters were normalized within each scenario. The differences, however, are shown by the mean person parameters for each scenario.

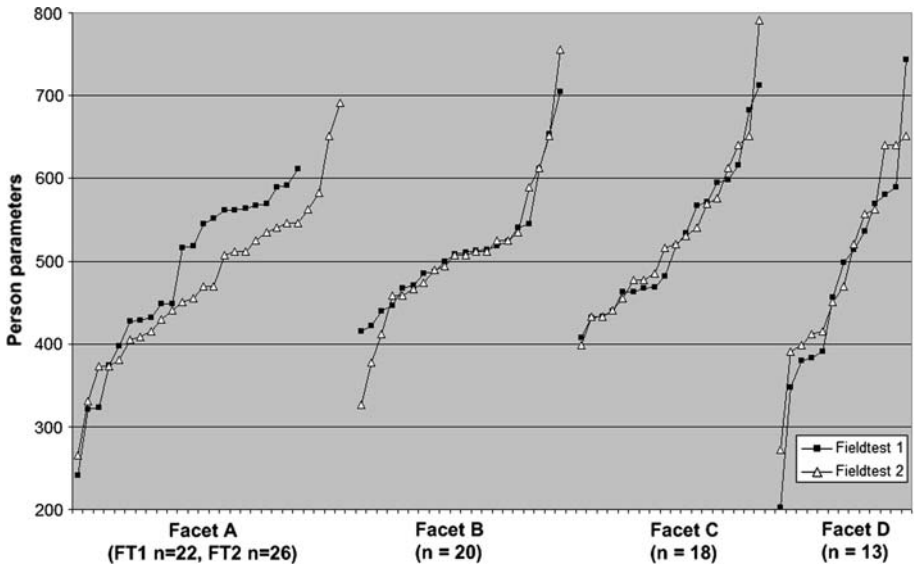
The items of the first scenario were the simplest. The mean person parameter (min–max) of scenario 1 was the lowest of all scenarios 394.90 (199.10–675.71) compared to the mean person parameters of scenario 2: 496.67 (133.07–763.04), scenario 3: 635.04 (372.73–973.34) and scenario 4: 473.39 (153.65–776.61).

Therefore, less competence (lower person parameters) is needed to solve the items of scenario 1. Furthermore, for curriculum evaluation with several measurements, scenarios can be selected by difficulties, starting with the simpler ones. Thus the instrument allows tailored testing. When testing is conducted with different scenarios it is necessary to adjust for different scenario difficulties in order to achieve comparability of person parameters. Therefore the mean value of the baseline sample has to be added, which is derived from the Weighted Likelihood Estimate (WLE) (Table 8). Maximum likelihood estimates (MLE) as well as weighted likelihood estimates (WLE), (Warm 1989) can be computed with WINMIRA. Warm’s WLE estimates have, as compared to the MLE estimates, two main advantages: First, their bias is smaller (Warm 1989; Hoijsink and Bomsma 1995), and second, they produce reasonable estimates even for extreme response patterns.

**Table 7** Second field test: BIC scores

Analyses	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Overall
Rasch (1 class)	2249.73	2451.22	1779.18	2662.31	8986.01
MiRa (2 classes)	2317.16	2556.69	1865.92	2742.71	9437.05
Hybrid Model	2272.97	2484.94	1837.40	2697.90	9162.78

Note. MiRa = Mixed Rasch Model



**Fig. 1** Item parameter sorted by categories of facets A–D

**Table 8** WLE and correction factors

	WLE <sup>a</sup>	Correction factors
Scenario 1	605.10	-105.10
Scenario 2	503.33	-3.33
Scenario 3	364.96	135.04
Scenario 4	526.61	-26.61

<sup>a</sup> Weighted likelihood estimate

### Construct validity

Person parameters are shown separately for university students and untrained secondary school students (Table 9). Effect size (Cohens *d*) was 4.33 (95% CI 3.51–5.16). University students are regarded as trained, since they attended at least one course at university that is similar to the described curriculum. Mean person parameters and effect size indicate a considerable difference between subsamples due to differences in attended school years and training at university of university students.

The second field test resulted in an improvement of overall reliability as well as reliabilities for the single scenarios. Reliability of the test for the Rasch model (ANOVA) was

**Table 9** Person parameter of students

	<i>n</i>	Person parameter <sup>a</sup>	<i>SD</i>
Trained students	13	716.14	53.74
Untrained students	94	470.11	59.63

<sup>a</sup> Values are means

0.91 and for the single scenarios 0.71 (scenario 1); 0.78 (scenario 2); 0.75 (scenario 3) and 0.80 (scenario 4). According to the Spearman Brown formula the reliabilities of the scenarios correspond to the magnitude of the overall reliability.

## Discussion

Evaluation of training in critical health literacy is challenging. We developed and validated the CHC test, an instrument with 72 items to measure critical health competences. Since the instrument was tested in untrained and trained samples it is applicable to curriculum evaluation as well as measurements of students' critical health literacy.

The CHC test has significant strengths. It is the first one to operationalize the construct critical health literacy. Results of model fit analyses have shown that the ordinary Rasch model is the best fitting model, implying scalability of all students. Furthermore, reliability of the instrument (0.91) is good. As an alternative to administer the entire test instrument, competences can be surveyed by applying only one of the four scenarios. This kind of a screening instrument is time saving, avoiding cognitive overload of retested persons. Moreover the opportunity to use different scenarios for multiple testing provides an important methodological advantage. Furthermore, the instrument can easily be administered.

The instrument and the instructions for raters will be made available in a separate manual. The design of the test using 4 parallel scenarios allows supplementing the test with further scenarios if necessary.

Our study has also limitations. Since the instrument is strongly linked to the concept of evidence-based medicine, other aspects of general health literacy, like competences addressing the health care system or communicative competences are not in the focus of this instrument.

Instruments, which allow valid measurements of critical health literacy independent of curriculum evaluation, are lacking. Recently, a review of literature identified 104 instruments. All instruments were developed for the evaluation of curricula of evidence-based medicine. Therefore, they are not applicable to be used in settings with lay people (Shaneyfelt et al. 2006).

## Implication for practice

Empowerment of patients and consumers aiming at informed decision-making requires health literate individuals. Schools are increasingly recognized as key settings for promoting health care activities. Since the instrument has been developed for samples of trained and untrained secondary school and university students, ranging from 15–42 years, the CHC test is applicable to this age group of adolescents and adults who passed grade 10. According to the stage of the CHC test evaluation, this instrument is applicable to students who have passed at least grade 10 secondary school. The instrument and the instructions for raters will be made available for free as from 2008 in a separate manual. It can be purchased via University of Hamburg, Unit of Health Sciences and Education.

A validation of the CHC test in groups of students and adults without grade 10 graduation, with and without vocational training, and also groups of health care professionals will enable the generalisability of our findings.

The psychometric properties, validity and feasibility of the CHC test allow the conclusion to recommend the use of the CHC test to measure critical health competences.

**Acknowledgements** We thank Anke Fuhlendorf for participating in the process of item construction and also for project administration. We also acknowledge the contribution made by all the participating students. We also thank the principals and teachers for supporting this project. In addition we thank the Institute for Quality and Efficiency in Health Care, Cologne, Germany, for funding this project.

## References

- Berger, B., Meyer, G., Steckelberg, A., & Mühlhauser, I. (2007). *Are laypersons able to learn and to use evidence-based medicine (Ebm) skills?* Paper presented at the 4th International Shared Decision Making Conference “Shared decision-making in diverse health care systems: Translating research into practice”, Freiburg, Germany.
- Deutsches PISA-Konsortium (2001). *PISA Basiskompetenzen von Schüler-innen und Schülern im internationalen Vergleich*. Opladen: Leske und Budrich.
- Dickersin, K., Braun, L., Mead, M., Millikan, R., Wu, A. M., Pietenpol, J., Troyan, S., Anderson, B., & Visco, F. (2001). Development and implementation of a science training course for breast cancer activists: Project LEAD (leadership, education and advocacy development). *Health Expectations*, 4, 213–220.
- European Network of Health Promoting Schools. (2004). The European Network of Health Promoting Schools – the alliance of education and health. Retrieved August 20, 2005: <http://www.euro.who.int/document/e62361.pdf>
- Hojtink, H., & Boomsma, A. (1995). On person parameter estimation in the dichotomous Rasch model. In G. H. Fischer & I. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications*. New York: Springer.
- Lynagh, M., Perkins, J., & Schofield, M. (2002). An evidence-based approach to health promoting schools. *The Journal of school health*, 72, 300–302.
- Max Planck Institute for Human Development (Ed.) (2000). *Third International Mathematics and Science Study – TIMSS*. Retrieved October 1, 2005 from Max Planck Institute for Human Development Web site: <http://www.timss.mpg.de/>
- Milne, R., & Oliver, S. (1996). Evidence-based consumer health information: Developing teaching in critical appraisal skills. *International Journal for Quality in Health Care*, 8, 439–445.
- Nutbeam, D. (2000). Health literacy as a public health goal: A challenge for contemporary health education and communication strategies into the 21st century. *Health Promotion International*, 15, 259–267.
- Rosenfeld, P., Salazar-Riera, N., & Vieir, D. (2002). Piloting an information literacy programme for staff nurses: Lessons learned. *Computers Informatics, Nursing*, 20, 236–241.
- Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion*. Bern: Huber.
- Rost, J., & von Davier, M. (1994). A conditional item fit index for Rasch models. *Applied Psychological Measurement*, 18(2), 171–182.
- Shaneyfelt, T., Baum, K. D., Bell, D., Feldstein, D., Houston, T. K., Kaatz, S., Whelan, C., & Green, M. (2006). Instruments for evaluating education in evidence-based practice: A systematic review. *JAMA: The Journal of the American Medical Association*, 296(9), 1116–1127.
- Simonds, S. K. (1977). Health education today: Issues and challenges. *The Journal of School Health*, 47(10), 584–593.
- Steckelberg, A., Hülfenhaus, C., & Mühlhauser, I. (2006) *Ebm@school: Ein Curriculum zur Kompetenzentwicklung von critical health literacy. Welche Interessen haben Schülerinnen und Schüler Allgemeinbildender Schulen?* Poster presented at the 7th Annual Meeting of the German Network for Evidence Based Medicine, Bochum.
- Taking health literacy seriously (Editorial) (2005). *Lancet*, 366, 95.
- von Davier, M. (2001). *WINMIRA: A software for analyses with a variety of discrete mixture distribution models for dichotomous and polytomous categorical data* (Germany).
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response models. *Psychometrika*, 54, 427–450.
- Yamamoto, K. (1989). *A Hybrid model of IRT and latent class models*. ETS Research Report (RR-89-41). Princeton, NJ: Educational Testing Service.