European Journal of Public Health, Vol. 18, No. 3, 294–299

doi:10.1093/eurpub/ckn001 Advance Access published on February 6, 2008

An international scoring system for self-reported health complaints in adolescents

Ulrike Ravens-Sieberer[1], Michael Erhart[1], Torbjorn Torsheim[2], Jorn Hetland[3], John Freeman[4], Mia Danielson[5], Christiane Thomas[1] and The HBSC Positive Health Group

Background: Aimed to develop a unitary scoring system for the 'Health Behaviour in school-aged Children' (HBSC) symptom checklist that would facilitate cross-national comparisons and interpretation. Rasch measurement analysis and investigation of differential item functioning (DIF) were conducted. **Methods:** Data were obtained from the 'WHO collaborative study HBSC 2001/2002'. A total of 162 305 students aged 11, 13 and 15 years from 35 European and North American Countries were surveyed. Unidimensionality of the items and local independence were tested using means of confirmatory factor analysis. DIF across countries, age groups and gender was investigated using a logistic regression procedure. Item and person parameters were estimated according to the Rating Scale Model (RSM). **Results:** All items proved to be unidimensional. One item displayed noticeable DIF across countries and was discarded. The remaining items were functioning equally across subgroups. The RSM analysis resulted in Rasch model conform item parameter estimation. Infit mean square values between 0.84 and 1.35 revealed acceptable item fit. **Conclusion:** The control of DIF enables comparable and unbiased assessment of subjective health complaints across countries, age groups and gender. A scoring algorithm could be developed which enables a cross-cultural comparable and interval-scaled assessment of subjective health complaints.

Keywords: children and adolescents, differential item functioning, HBSC-Health Survey, international comparison, Rasch analysis, subjective health complaints

Introduction

Subjective health complaints like headache, backache, feeling low etc. are common in adolescence[1–3] and tend to occur in cluster rather than as single symptoms.[1,4,5] In few instances these symptoms are related to a defined diagnosis or disease.[6] Previous research has found that these symptoms tend to increase with age and are more prevalent in girls.[2]

With trends of globalization in public health, assessment of cross-national differences in subjective health becomes increasingly important.[7–10] Cultures might not only differ in the frequency of psychosomatic health complaints, but also in the specific complaints expressed and may be in the exact meaning of the concept. The comparison and interpretation of previous studies is limited by methodological differences regarding e.g. the definition of symptoms and time frame of reporting. A previous study found the pattern of adolescents subjective health complaints to be consistent across countries, although the prevalence decreased from Finnish to Scottish, Polish and Norway students.[9] However another study in Scandinavians aged 15 and older observed no clear differences between Sweden, Denmark, Norway and Finland examining any complaints at all. But with regards to substantial complaints, the Swedes had the highest reports, while the Finnish had the fewest reports.[11]

Such investigation calls for measurement instruments that can be applied and compared across populations, languages and demographic subgroups.[12,13] A self-report rating scale that has been used with such a comparative perspective is the Health Behaviour in school-aged Children symptom checklist (HBSC-SCL).[8,14,15] The HBSC-SCL assesses the occurrence of eight common health complaints. The HBSC-SCL has been translated and used in more than 35 countries,[16] but a cross-culturally uniform scoring system is still lacking. The present study presents a scoring algorithm on a sample of 162 305 students from 35 European and North American countries that would ensure such a scoring.

A basic requirement for any scoring system of health complaints is interpretability and consistency across samples. In several previous studies, the scoring of the HBSC-SCL has been based on raw summation of item scores. This scoring system has achieved high reliability,[14] but enables measurement at ordinal scale level only. As an alternative, scoring based on the Rasch model can provide measurement on interval scaled level, which is desirable for epidemiological cross-cultural comparison, trend-studies over time or studies on the influence of environmental factors on health. Rasch models belong to the so-called item response theory (IRT). IRT models assume a test-persons response to an item can be explained by his/her trait parameter value (θ) which can be considered as his/her position on a latent trait continuum and the position of thresholds (δ) between the item answer categories on the same latent trait continuum that is the position where neighbouring answer categories are chosen with the same probability—below δ the lower answer category is more likely above δ the higher answer category is more likely to be chosen. Some IRT models encompass additional model parameters as well. In Rasch model, the probability of choosing an answer category is modelled by a logistic function of the difference between θ and δ. Rasch-based scaling has several properties that are attractive, including a psychologically meaningful model of measurement and statistics of individual

[1] School of Public Health, WHO Collaborating Center for Child and Adolescent Health Promotion; University of Bielefeld, Germany
[2] University of Bergen, Norway
[3] Research Centre for Health Promotion, Bergen, Norway
[4] Queen's University, Kingston, Ontario, Canada
[5] National Institute of Public Health, Stockholm, Sweden

Correspondence: Prof. Dr Ulrike Ravens-Sieberer, School of Public Health, Director WHO Collaborating Center for Child and Adolescent Health Promotion, University of Bielefeld, PB 10 01 31, D-33501 Bielefeld, Germany, tel: +49 521 106 6245, fax: +49 521 106 6433, e-mail: U.Ravens-Sieberer@uni-bielefeld.de

measurement error. Unidimensional constructs can be identified and biased or weak indicators can be detected. If the Rasch model fits the data well the sum of the item-scores represents a sufficient statistic for the response to all items. Examining the single item scores would then not add any crucial psychometric information to the measurement.[17,18,19] Applying the HBSC-SCL as a short screening instrument requires setting up thresholds for a noticeable outcome. Rasch-based scaling provides the possibility of linking scores with the meaningful item-content[17] thus helping to define cut-off points.

For a scale to be scored according to the Rasch model, the covariation between the items of the scale has to satisfy the relatively stringent assumptions of *local independence* and *unidimensionality* (the item score should not be related to any other item score expect through the contribution of both to the scale score).[17,18] Two previous studies of the dimensionality of the HBSC-SCL with confirmatory factor analysis (CFA) revealed adequate fit for a one-factor model. A correlated two-factor model (physical and psychological) was clearly superior.[14,15] However, the correlation between these two factors was extremely high (0.80–0.82). Thus, in the present study, our expectation was that the HBSC-SCL would show sufficient unidimensionality.

Multidimensionality may result in poor Rasch item fit statistics.[20] A way to prevent this is to apply explicit tests of unidimensionality, which might be more sensitive and specific than the CFA analyses[14,15] which e.g. can fail in the presence of differing item difficulties.[18,21,22]

Rasch modelling of the HBSC symptom checklist was applied in a recent study.[23] The partial credit model (PCM) was examined in a sample of Swedish adolescents, using the HBSC-SCL. Reversed thresholds—that is at least one answer category is at no position of the latent trait most likely to be chosen (which does not necessarily mean that the answer categories operate in a reversed order)—were found for three items. As a solution, the authors suggested discarding the three mentioned items. The five remaining items showed item characteristic curves that were consistent with the PCM.

This study[23] highlights some of the challenges in applying the Rasch model to existing rating instruments, including the problem of how to deal with violations of the Rasch model assumptions. When reversed thresholds are present, one alternative to discarding items could be to collapse response categories.[18,24] A problem with collapsing response categories is the potential loss of psychometric information. A second feasible alternative would be to force the threshold estimates being in the right order and to test whether response patterns are consistent with such variants of the Rasch model.

For scales with a uniform set of response categories across items, such as the HBSC-SCL, the Rasch Rating scale model (RSM)[25] may be a good alternative to the PCM. The assumption made in the RSM is that items differ in their location on the latent trait, but that the distance between thresholds and the order of category thresholds are the same across items[25]

The aims of this study were to investigate:

(1) If the actual response behaviour of the tested adolescents on the HBSC-items can be reasonably explained by the assumption of ordered thresholds and a unidimensional latent trait continuum.
(2) If the items are functioning in the same way across countries, age groups and gender.
(3) If a new scoring algorithm could be established.
(4) The results of the new scoring algorithm in terms of demographic differences and the association with health status.
(5) The results of the new scoring in terms of cross-national variation and consistency.

Methods

Sample and data collection

Data were obtained from the large cross-national WHO collaborative study ‘Health Behaviour in School-aged Children’ 2001/2002.[26] A total of 162 305 students from 35 countries and regions participated, 51.7% were females, 33.6% were aged 10–12, 34.6% were aged 12–14, 31.8% were aged 14–16.

The sample was selected through a complex multistage sampling procedure. The primary sampling unit was school-class, with self- selection of students. More detailed information about the sample and the sampling frame can be obtained elsewhere.[26,27] The multistage sampling makes it difficult to compute a single response rate. The available documentation suggest that for a majority of countries, the response rate at the level of school was above 80% with additional dropout at the student-level ranging from 2.4 to 26.0%. Combining all available information the weighted response rate was 75% (authors calculation).

Instrument and variables

The eight items of the HBSC-Symptom checklist ask about how often in the last 6 months the children and adolescents have suffered from the following complaints: headache, stomach ache, backache, feeling low, irritable or bad tempered, feeling nervous, sleeping difficulties, dizziness. The children and adolescents can respond to these questions by choosing one of the following five answer categories: ‘rarely or never’, ‘about every month’, ‘about every week’, ‘more than once a week’ and ‘about every day’.[27] Additional information used for the analyses included the age, gender and self-reported general health status of the respondents.

Statistical analysis

Subjects with missing values in any variable (2.9% of all cases) were omitted from the analysis. A confirmatory factor-analysis was computed (Mplus) by specifying a one-factor-structural equation model across all items using polychoric correlations. Items with loading above 0.4 remained in the item pool. To assess local dependence, the common factor was partialized out of the items and the item residuals were correlated with each other.[28] If the residuals of two or more items correlated above 0.2 with each other, one or more of the involved items were eliminated.[29]

Students with the same level of trait should respond similarly to an item, regardless of their culture, age or gender. Differential item functioning (DIF) occurs when people at the same level of trait but from e.g. different countries respond differently. To examine DIF, we used the logistic regression approach described by Zumbo:[13] Every item serves as the dependent variable in hierarchical ordinal regression models. The goodness of fit of a logistic regression model with the total score being the only covariate was compared with the goodness of fit of a model where the total score, the group and the group × total score interaction were the covariates. The significance of the χ^2-changes, as well as the change in the Nagelkerke pseudo-R^2, was investigated. While the first value tells about significant uniform (different locations on the latent trait), non-uniform (different slopes) and absolute DIF, the second statistic gives an impression about the effect size of the DIF. A pseudo-R^2 change of 0.035 has been suggested as a criterion for practically meaningful DIF.[30] We thus set the threshold for a tolerable DIF effect to an R^2 change of 0.035 but acknowledged that other authors proposed quite less restrictive thresholds of 0.07.[13] The analyses were carried out using the PLUM procedure of SPSS.

Table 1 Differential item functioning across countries, age and gender – identified with logistic regression procedure

	DIF country[a]		DIF age[b]		DIF gender[c]	
Item	R^2**-change**[d]	P[d]	R^2**-change**[c]	P[d]	R^2**-change**	P[d]
Headache	0.013	<0.001	<0.001	<0.001	0.003	<0.001
Stomach ache	0.019	<0.001	0.004	<0.001	0.014	<0.001
Backache	0.014	<0.001	0.012	<0.001	0.003	<0.001
Feeling low	0.036	<0.001	0.003	<0.001	0.001	<0.001
Irritable bad tempered	0.023	<0.001	0.002	<0.001	0.002	<0.001
Nervous	0.037	<0.001	0.001	<0.001	0.001	<0.001
Sleeping difficulties	0.045	<0.001	0.003	<0.001	0.002	<0.001
Dizziness	0.015	<0.001	<0.001	<0.001	0.004	<0.001

a: Random 5% of the overall sample versus non-selected single country-samples

b: 15-year olds versus 11-year olds and versus 13-year olds

c: Girls versus boys

d: Transgression probability of the difference in the –2loglikelihood χ^2-value; and Difference in the Nagelkerke R^2-value between Ordinal Logistic Regression 'base' model (item regressed on total score) and 'uniform/non-uniform DIF' model (item regressed on total score, group, total score × group). The group is defined by country, age or gender

Table 2 Item fit and parameter estimation according to the rating scale model (extra-conditional joint maximum likelihood estimation)

	Infit msq[a]	**Location**[b,c]	**Step1**[b,d]	**(SE)**	**Step2**[b,d]	**(SE)**	**Step3**[b,d]	**(SE)**	**Step4**[b,d]	**(SE)**
Headache	1.02	48.99	44.1	(0.02)	47.8	(0.03)	48.5	(0.03)	55.5	(0.05)
Stomach ache	0.97	51.85	47.0	(0.02)	50.7	(0.03)	51.4	(0.03)	58.3	(0.05)
Backache	1.35	53.49	48.6	(0.02)	52.3	(0.03)	53.0	(0.03)	59.9	(0.05)
Feeling low	0.92	49.12	44.3	(0.02)	48	(0.03)	48.7	(0.03)	55.6	(0.05)
Irritable-bad temper	0.84	45.08	40.2	(0.02)	43.9	(0.03)	44.6	(0.03)	51.5	(0.05)
Nervous	0.96	46.00	41.1	(0.02)	44.9	(0.03)	45.5	(0.03)	52.5	(0.05)
Dizziness	1.24	55.47	50.6	(0.02)	54.3	(0.03)	55.0	(0.03)	61.9	(0.05)

a: Infit mean square values between 0.7 and 1.3 indicate good fit to the Rasch model

b: Item-parameter-based metric (sample mean = 38.5; SD = 12.1)

c: Mean of step-parameter

d: Steps = threshold parameter (location on latent trait continuum were neighbouring answer categories are chosen with equal probability)

The estimation of the Rasch RSM parameters was done using the WINSTEPS program joint maximum likelihood procedure.[31] In this procedure, the 'person-parameter' of every single respondent is estimated simultaneously with the item parameter. The disadvantages of this procedure are that it may lead to inconsistent estimations especially in large samples and short scales of <10 items.[18,19] To correct for this potential bias we used the extra-conditional method (XCON, XMLE).[32]

In the present study, data fit was indicated by the infit mean square statistic, which is based on the residuals between the empirical and the theoretical expected item scores.[18] In line with conventional criteria,[24] a well fitting item would be expected to have an infit mean square between 0.7 and 1.3.

Results

Unidimensionality and local independence

Unidimensionality and local independence were tested using CFA. In the specified one-dimensional model, the item-loadings ranged between 0.52 (Backache) and 0.70 (Feeling low; Irritable-bad temper). The common factor accounted for 39.9% of the variance in the items. In the next step of analysis, the common factor was partialized out of the items, and the item residuals were correlated with each other. Item residual correlation ranged between 0.01 and 0.12. These values are clearly below the conventional threshold of 0.2, which has been used to indicate a secondary factor or violation of local independence. Running a conventional exploratory factor analysis on the data resulted in only one factor having an eigenvalue >1 (data not shown).

Differential item functioning

Table 1 shows the results for tests of DIF. For the logistic regression DIF procedure examining differences across country, a random sample of 5% of the adolescents from each country was selected as the reference group and contrasted with the 35 country samples (without the selected 5%). The item 'sleeping difficulties' displayed sizeable country DIF (R^2-change = 0.045) and thus was excluded from further analysis. All other items remained in the scale because their magnitude of country DIF was below or only slightly above the defined criterion 0.035. None of the eight original items displayed noticeable DIF across age groups and gender.

Rasch modelling

The results of fitting the RSM to the HBSC symptom checklist are shown in table 2. The table includes item parameters estimated from the joint maximum likelihood estimation procedure of WINSTEPS. Parameters were calibrated to a common metric defined by the mean of the item-parameters (set to 50) and their SD (set to 10). It can be seen from the table that the locations of the items differed. The item 'Irritable' had the lowest location at 45.0, meaning that responses to this item tap information at a comparatively lower level of the trait than the other items. In contrast, the item 'dizziness' had a location at 55.6, providing maximum information about scores above the average of the items.

The table also provides the thresholds for steps between response categories. It can be seen that increasing thresholds (steps between categories) relate to increasing trait levels. The difference between step 2 and 3 was only 0.7 units, indicating that the middle answer category 'about every week' has little discriminatory power. According to the model, those respondents who reported having symptoms 'about every week' did not differ appreciably from those responding that they had experienced symptoms 'about every month' or from those who reported symptoms 'more than once a week'.

All items fit the data well according to the RSM. Infit mean square values ranged between 0.84 and 1.35. Using a fairly conservative criterion, only the item 'backache' displayed slight misfit according to the recommendations of some researchers, with the actual infit mean square value of 1.35 indicating lower discriminatory power than theoretically expected. The person separation index indicating overall reliability of the measurement was 0.77, (Cronbach's alpha = 0.78).

The new Rasch scoring algorithm for the HBSC-SCL encompasses summing up the seven item scores. The sum-scores are then non-linearly transformed into Rasch scores (Sample mean = 38.5; SD = 12.1).

Demographic differences

Table 3 shows the mean Rasch score for demographic subgroups, split by gender. For boys and for girls, scores increased with increasing age. A 11 year old girl would on an average have a score of 37.3, whereas a 15 year old girl would be expected to have a score of 43.1, indicating a major shift from early to mid-adolescence. Across SES subgroups, there were only minor differences. Young people from intact families had lower scores than other family structures.

Association with self-rated health

Though the self-rated health cannot be viewed as a validity criterion for a scoring algorithm of subjective health complaints, the association between both was assessed because from a theoretical point of view a sizeable relationship could be expected. Table 3 shows the mean Rasch score for respondents with different health rating. Boys (girls) rating their health as 'excellent' on average have a score of 33.2 (35.7) whereas boys (girls) with 'poor' health rating on average have a score of 48.5 (52.1), indicating a noticeable increase of nearly 1.5 SD.

Cross-national variation and consistency

To be used in a cross-national setting, the symptom checklist should be sensitive to true cross-national differences. To examine the amount of cross-national variation, a series of mixed models with random effects at the country level was computed, using the software MlwiN.[33] It can be seen from table 4 that the random cross-national variation in the HBSC-SCL Rasch scores was statistically significant across all age groups and gender. The variance partitioning coefficient was highly consistent, ranging from 0.044 to 0.063 across age groups and gender, indicating a high degree of consistency in the overall magnitude of random effects at the country level.

With regards to country average level of subjective health complaints several cluster of countries emerged: the mid and eastern Mediterranean countries Israel = 44.0, Italy = 43.9, Malta = 41.0, Greece = 40.2 displayed high levels as well as the eastern Scandinavian countries Sweden = 41.7 and Finland = 40.3; whereas the western Scandinavian countries Norway = 37.7 and Denmark = 36.7 displayed lower scores. The Baltic countries Estonia = 38.7, Lithuvia = 37.9 and Latvia = 37.5 are in the mid range. The former Yugoslavian countries Croatia = 36.8, Slovenia = 35.3 and Macedonia = 35.2 are among the countries with low levels.

To examine the consistency of the cross-national differences in HBSC-Rasch scores, the correlation between the country mean score for girls and for boys was computed.

Table 3 Mean rasch scores for sociodemographic subgroups and different self rated health

	Boys		Girls	
	Mean[a]	95% CI	Mean[a]	95% CI
Age group				
11 year olds	35.26	(35.04–35.48)	37.33	(37.12–37.53)
13 year olds	36.66	(36.46–36.85)	40.74	(40.56–40.91)
15 year olds	37.72	(37.52–37.91)	43.13	(42.98–43.29)
SES[b]				
Low	37.05	(36.84–37.27)	41.28	(41.10–41.45)
Medium	36.13	(35.97–36.29)	40.06	(39.91–40.21)
High	36.61	(36.43–36.79)	39.96	(39.78–40.14)
Family structure				
Intact	36.15	(36.02–36.28)	39.96	(39.84–40.09)
Stepfamily	38.13	(37.79–38.46)	42.38	(42.10–42.66)
Single mother	37.38	(37.10–37.67)	41.50	(41.25–41.74)
Single father	38.19	(37.51–38.88)	42.44	(41.78–43.09)
Other care	38.15	(37.03–39.28)	43.28	(42.45–44.10)
Self-rated health				
Poor	48.51	(47.65–49.37)	52.10	(51.58–52.62)
Fair	41.60	(41.36–41.85)	45.95	(45.78–46.12)
Good	36.98	(36.86–37.10)	40.42	(40.32–40.53)
Excellent	33.23	(33.08–33.38)	35.67	(35.51–35.84)

a: Item parameter based metric (sample mean = 38.5; SD = 12.1)

b: Assessed via Family affluence scale (low, medium and high familial affluence)

Table 4 Random intercept model with random effects for countries

	Intercept[a,b] (SE)	Uj[c] (SE)	Eij[d] (SE)	VPC[e]	95% CI country
11 year olds					
Boys	35.05 (0.54)	10.05 (2.45)	148.53 (1.31)	0.063	(28.84–41.26)
Girls	37.19 (0.46)	7.33 (1.80)	139.27 (1.21)	0.050	(31.88–42.49)
13 year olds					
Boys	36.47 (0.05)	7.46 (1.83)	133.04 (1.12)	0.053	(31.12–41.82)
Girls	40.53 (0.40)	5.44 (1.33)	101.60 (0.86)	0.051	(35.96–45.10)
15 year olds					
Boys	37.53 (0.05)	5.56 (1.14)	120.08 (1.12)	0.044	(32.91–42.15)
Girls	42.85 (0.37)	4.66 (1.14)	83.35 (0.73)	0.053	(38.62–47.08)

a: Item parameter based metric (sample mean = 38.5; SD = 12.1)

b: Regression constant = mean Rasch-score

c: Amount of variance accounted by between country differences in the regression constant (mean Rasch scores)

d: Amount of variance accounted for by individual differences in Rasch scores

e: Proportion of overall variance in Rasch scores accounted for by between country differences in the regression constant (mean Rasch scores)

Countries' mean scores for boys were strongly associated with their mean scores for girls. The Pearson correlations between countries mean scores for 11, 13 and 15 years old boys and girls were 0.93, 0.92 and 0.90, respectively, revealing a high consistency in country Rasch scores for independent subgroups.

Discussion

The main objective of the present article was to develop a scoring system for the HBSC symptom-checklist that would enable fair comparison and interpretation across countries. The main outcome was a unidimensional scoring algorithm based on seven of the eight original items. Previous studies on the HBSC-SCL suggested a two-factor solution (physical and psychological). However our analyses indicated unidimensionality of the HBSC-SCL items. The assumption of a unidimensional latent trait showed to be sufficient to explain the actual response behaviour of the respondents. As the HBSC symptom checklist actually comprises of 8 items only, we would not recommend splitting it into a 5 item and a 3 item domain to be scored using Rasch modelling. Such a model would be less reliable and eventually would show poor person fit. Most important—as the unidimensional Rasch-model already fits the data well—a two-factor model would not add crucial psychometric information to the measurement.

Compared with the more commonly used summed raw score, the Rasch based scaling has major advantages in terms of interpreting severity and content of scale scores. From the results it can be stated that frequent dizziness and backache, for example, are indicators of severe physical health complaints, whereas mild or moderate health complaints are most likely characterized by the absence of all complaints except feeling nervous or being bad tempered. Adolescents with a parameter value of, for example, 60 (Percentile 99) are most likely to be suffering every day from all of the seven complaints except dizziness and backache—which they are most likely to be suffering 'more than once a week' (but not 'every day'). Adolescents scoring at the medium score (40) are most likely to be suffering at least 'about every month' from being 'bad tempered' or 'feeling-nervous' but are most likely to be free ('rarely or never') from any of the other health complaints.

It is important to note that the Rasch analysis revealed a higher measurement precision on the higher end of the trait continuum. This finding can be seen as a valuable information for epidemiological screening, as it is more important to distinguish between medium and high severity than between medium and low severity.

The test of DIF revealed that the item 'sleeping difficulties' worked differently across the different countries, indicating that this item may introduce bias in cross-cultural comparisons. To achieve the essential objective of comparability across countries, a decision to discard this item was made. The remaining seven items enable a cross-culturally comparable and unbiased assessment of subjective health complaints for boys and girls between 11 and 15 years although two of them slightly exceeded the a priori set threshold. Yet this deviation still might be attributable to chance.

According to the a priori defined criterions all remaining items met the important Rasch model assumptions of locally independent item responses. All items displayed reasonable fit statistics. Using the one-dimensional score of the Rasch person parameters would not lead to loose crucial psychometric information. The assumption of ordered thresholds could explain the empirical test data well, though the small distance between thresholds 2 and 3 indicated weaknesses in the response choices: 'about every week' is likely too close to its neighbouring answer categories.

The scoring algorithm obtained in the present study revealed a consistent pattern of cross-national differences with historically/cultural similar countries having similar scores. It enables a cross-cultural comparable interval scaled assessment of subjective health complaints in school-aged children, which is unbiased regarding age and gender. Further (qualitative) studies might focus on investigating if the HBSC-SCL is sensitive for gender specific aspects of health complaints. Qualitative analysis of item content could be applied to set up threshold for distinguish between negligible and noticeable subjective health complaints.

The new scoring algorithm is available as an SPSS syntax, however syntaxes for other statistical software are also possible. The scoring could be used for international comparisons using the HBSC-SCL.

Acknowledgements

The World Health Organization-Health Behaviour in School-aged Children Survey is a WHO/EURO collaborative study. The authors would like to thank the international coordinator of the 2001/2002 study Candace Currie, University of Edinburgh, Scotland; and the data bank manager Oddrun Samdal, University of Bergen, Norway; and the HBSC Positive Health Group: Ulrike Ravens-Sieberer, Torbjørn Torsheim, Bogdana Alexandrova, Antony Morgan, Kädi Lepp, Raili Välimäa, Céline Vignes, Cornelia Haehne, Jennifer Nickel, Nadine Helle, Christina Schnohr, Gyöngyi Kökönyei, Franco Cavallo, Inese Gobina, Wilma Vollebergh, Saskia van Dorsselaer, Jørn Hetland, Joanna Mazur, Tania Gaspar, Viorel Mih, Aurora Szentagotai, Eva Kallay, Helena Jericek, Eva Stergar, Vesna Pucelj, Pilar Ramos, Mia Danielson, Mujgan Alikasifoglu, Ethem Erginoz, Mary Overpeck.

Conflict of interest: None declared.

Key points

- A scoring system for the 'Health Behaviour in school-aged Children' (HBSC) symptom checklist that would facilitate cross-national comparisons was developed.
- Data from 35 European and North American Countries were obtained from the 'WHO collaborative study HBSC 2001/2002'.
- Rasch measurement analysis and investigation of DIF using logistic regression methods showed one out of eight items to display sizeable DIF which thus was discarded.
- Unidimensionality of the items was proved using means of CFA. Rasch model analysis resulted in conform item parameter estimation and acceptable item fit.
- A scoring algorithm could be developed which enables a cross-cultural comparable and interval-scaled assessment of subjective health complaints to be included in further HBSC or other epidemiological studies.

References

1 Garralda ME. Somatisation in children. *J Child Psychol Psychiatry* 1996;37:13–33.

2 King A, Wold B, Tudor-Smith C, Harel Y. The health of youth. A cross national survey. *WHO Regional Publications. European Series*;69: Kopenhagen: WHO Regional Office for Europe.

3 Mikkelsson M, Salminen J, Kautiainen H. Nonspecific muskuloskeletal pain in pre-adolescents. Prevalence and 1-year persistence. *Pain* 1997;73:29–35.

4 Knishkowy B, Palti H, Tima C, et al. Symptom clusters among young adolescents. *Adolescence* 1995;30:351–62.

5 Alfven G. The covariation of common psychosomatic symptoms among children from socio-economically differing residential areas: an epidemiological study. *Acta Paediatr* 1993;82:484–7.

6 Garralda ME. A selective review of child psychiatric syndromes with a somatic presentation. *Br J Psychiatry* 1992;161:759–73.

7 Kirmayer LJ, Young A. Culture and somatization: Clinical, epidemiological, and ethnographic perspectives. *Psychosom Med* 1998;60:420–30.

8 Crijnen AAM, Achenbach TM, Verhulst FC. Problems reported by parents of children in multiple cultures: the child behavior checklist syndrome constructs. *Am J Psychiatry* 1999;156:569–74.

9 Haugland S, Wold B, Stevenson J, et al. Subjective health complaints in adolescence - A cross-national comparison of prevalence and dimensionality. *Eur J Public Health* 2001;11:4–10.

10 Gureje O, Simon GE, Ustun TB, Goldberg DP. Somatization in cross-cultural perspective: A World Health Organization study in primary care. *Am J Psychiatry* 1997;154:989–95.

11 Erjksen HR, Svendsrod R, Ursin G, Ursin H. Prevalence of subjective health complaints in the Nordic European countries in 1993. *Eur J Public Health* 1998;8:294–8.

12 Camilli G, Shepard LA. *Methods for identifying biased test items*. Thousand Oaks, California: Sage, 1994.

13 Zumbo BD. *A Handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores*. Ottawa: Directorate of Human Resources, Research and Evaluation, Department of National Defense, 1999.

14 Haugland S, Wold B. Subjective health complaints in adolescence-Reliability and validity of survey methods. *J Adolesc* 2001;24:611–24.

15 Hetland J, Torsheim T, Aaro LE. Subjective health complaints in adolescence: dimensional structure and variation across gender and age. *Scand J Public Health* 2002;30:223–30.

16 Torsheim T, Välimaa R, Danielson M. Health and well-being. In: Currie C, Roberts C, Morgan A, Smith R, Settertobulte W, Samdal O, et al, editors. *Young people's health in context*. Copenhagen: World Health Organization, 2004: 55–62.

17 Hambleton RK, Swaminathan H, Rogers HJ. *Fundamentals of item response theory*. Newbury Park, California: Sage, 1991.

18 Embretson SE, Reise S. *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers, 2000.

19 Rost J. *Lehrbuch Testtheorie und Testkonstruktion*, 2nd edn. Bern: Huber, 2004.

20 Smith EV Jr. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *J Appl Meas* 2002;3:205–31.

21 Hattie JA. An empirical study of various indices for determing unidimensionality. *Multivariate Behav Res* 1984;19:49–78.

22 Nunnally JC, Bernstein IR. *Psychometric theory*, 3rd edn. New York: McCraw-Hill, 1994.

23 Hagquist C, Andrich D. Measuring subjective health among adolescents in Sweden: a Rasch-analysis of the HBSC Instrument. *Soc Indic Res* 2004;68:201–20.

24 Bond TG, Fox CM. *Applying the Rasch model: fundamental measurement in the human sciences*. Mahwah, NJ: L. Erlbaum, 2001.

25 Andrich D. A rating formulation for ordered response categories. *Psychometrika* 1978;43:561–73.

26 Currie C, Roberts C, Morgan A, et al., editors. *Young People' Health in context. Health Behaviour in School-aged Children (HBSC) study: international report from the 2001/2002 survey*. Copenhagen: World Health Organization, 2004.

27 Currie C, Samdal O, Boyce W, Smith R, editors. *Health behaviour in school-aged children: a World Health Organization cross-national study (HBSC). Research protocol for the 2001/2002 survey*. Edinburgh: University of Edinburgh, 2001.

28 Yen WM. Scaling performance assessments - strategies for managing local item dependence. *J Educ Meas* 1993;30:187–213.

29 Bjorner JB, Kosinski M, Ware JE Jr. Calibration of an item pool for assessing the burden of headaches: An application of item response theory to the Headache Impact Test (HIT-super(TM)). *Qual Life Res* 2003;12:913–33.

30 Jodoin MG, Gierl MJ. Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Appl Meas Educ* 2001;14:329–49.

31 Linacre JM, Wright BD. *A user's guide to WINSTEPS, BIGSTEPS, MINISTEP–Rasch–model Computer Programs*. Chicago: MESA Press, 1999.

32 Linacre JM. Extra-conditional (XCON, XMLE) algorithm. *Rasch Measure Transac* 1989;3:47–8.

33 MLwiN [program]. 1.10.0007 version. London: Multilevel Models Project, 2001.

Received 31 May 2007, accepted 24 December 2007