<https://helda.helsinki.fi>

Differential Item Functioning in Trend Analyses of Adolescent Mental Health - Illustrative Examples Using HBSC-Data from Finland

Hagquist, Curt

2017-09

Hagquist , C , Valimaa , R , Simonsen , N & Suominen , S 2017 , ' Differential Item Functioning in Trend Analyses of Adolescent Mental Health - Illustrative Examples Using HBSC-Data from Finland ' , Child Indicators Research , vol. 10 , no. 3 , pp. 673-691 . <https://doi.org/10.1007/s12187-016-9397-8>

<http://hdl.handle.net/10138/240354>

<https://doi.org/10.1007/s12187-016-9397-8>

unspecified

publishedVersion

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.

Differential Item Functioning in Trend Analyses of Adolescent Mental Health – Illustrative Examples Using HBSC-Data from Finland

Curt Hagquist¹ · Raili Välimaa² · Nina Simonsen³ · Sakari Suominen⁴

Accepted: 8 May 2016 / Published online: 30 May 2016
© Springer Science+Business Media Dordrecht 2016

Abstract Although there is an increasing focus on trend analyses of adolescent mental health, yet too little attention is paid to the methodological challenges and pitfalls inherent in this type of analyses. The purpose of the study is to analyse the psychometric properties of a Finnish instrument on psychosomatic problems, with a major focus on Differential Item Functioning (DIF) across time. Questionnaire data collected in 1994, 1998, 2006 and 2014 among Finnish schoolchildren in grade 9 (15-year-olds) as part of the Health Behaviour in School-aged Children (HBSC) study were utilised. The polytomous Rasch model was used to examine the psychometric properties of a composite measure of psychosomatic problems. The results clearly indicate that the composite measure on psychosomatic problems consisting of nine items does not work invariantly over time. In particular, the item depressed shows DIF across years of investigations. This item works quite differently at the first year of investigation compared to the last year showing higher expected values 2014 (=less frequent problems) than 1994. This DIF affects the person measure of change in psychosomatic problems between 1994 and 2014. Resolving the item depressed for year of investigation DIF, or removing it, increases the difference in person mean values between the two years, implying increasing psychosomatic problems over time. Since the DIF affects the trend results, different options to address the problems need to be considered. Removing the item depressed would bring the Finnish measure of psychosomatic

✉ Curt Hagquist
curt.hagquist@kau.se

¹ Centre for Research on Child and Adolescent Mental Health, Karlstad University, SE-651 88 Karlstad, Sweden

² Department of Health Sciences, University of Jyväskylä, Jyväskylä, Finland

³ Department of Public Health, University of Helsinki and Folkhälsan Research Center, Helsinki, Finland

⁴ University of Turku, Finland and University of Skövde, Skövde, Sweden

problems in better accordance with the content of the questions on psychosomatic problems in the international HBSC protocol in which the item depressed is not included.

Keywords Adolescents · HBSC · Mental health · Psychosomatic problems · Rasch · Trend analyses

1 Introduction

Adolescent mental health is a cause for concern worldwide. A recent systematic review of the international literature confirmed an increase in internalizing problems in the 21st century among girls while the results for boys were inconsistent (Bor et al. 2014). Given the interest for trends in adolescent mental health, surprisingly little attention has been paid to the methodological challenges and pitfalls inherent in these kinds of analyses. In fact, retaining the questions in their original format and using standardised data collection procedures still seems to be the major criteria considered in trend studies. In cross country comparisons of adolescent mental health, the focus is mainly on the use of procedures for translation and adaptation of questionnaires from one language/culture to another. There are, however, examples of recent studies conducted with a focus on cross country comparisons of the operating characteristics of the questions that the analyses are based on (Cieza et al. 2015).

Requirements for measurement invariance are commonly not brought up to front. In trend studies based on data from the Health Behaviour in School-aged Children (HBSC) study analyses of Differential Item Functioning (DIF) across years of investigations are uncommon, i.e. analyses of how individual items are functioning over time. Similarly, only a few analyses of DIF across countries of the HBSC-checklist on psychosomatic problems have been conducted (Ravens-Sieberer et al. 2008). In a recent paper based on international data from the HBSC study, time trends in psychosomatic problems in 35 countries were reported, without information about the operating characteristics of the instrument across time and between countries (Ottová-Jordan et al. 2015).

An illustrative example of the challenges in making cross-country comparisons can be found in the field of educational testing, where a controversial debate about cross-country DIF took place in the wake of the last survey of PISA, The Programme for International Student Assessment (Kreiner and Christensen 2014).

In the present paper we are addressing methodological challenges in trend analyses, in particular issues concerning invariant functioning of a measurement instrument across years of investigations. While this may be the most decisive criterion in trend analyses, in practice it is often neglected. We make use of data on psychosomatic problems collected repeatedly with questionnaires in Finland, not only in Finnish but also in Swedish, which also is an official language spoken by about five per cent of the population (Statistics Finland 2014), mainly in the southern and western parts of the country.

The purpose of the paper is to analyse the psychometric properties of a composite measure of psychosomatic problems, with a focus on the functioning of the items across years of investigations. The major research questions are a) if the instrument

works invariantly over time, i.e. if there is any evidence of DIF across years of investigations; b) to what extent potential DIF may affect the trend patterns.

2 Methods

Data were collected as part of the *Health Behavior in School-aged Children* (HBSC) study conducted in collaboration with the World Health Organization since the 1980s. The HBSC-study is targeting students in grades 5, 7 and 9, and now includes more than 40 countries (Currie et al. 2012). Data were collected with questionnaires which were completed anonymously in school classrooms. Participation was voluntary. Data collected at four points in time among Finnish schoolchildren are used. In the analyses conducted in the present study only data from grade 9 students are used, except for some DIF-analyses which tentatively also were conducted among the entire sample of grade 5, 7 and 9 students.

The total number of participants in grade 9 at each year of investigation was, 1994: 1716 (1194/522), 1998: 2145 (1545/600), 2006: 2168 (1670/498), 2014: 2476 (1935/541). The numbers within parentheses refer to the number of participants in Finnish and Swedish speaking schools respectively. It should be noted that all students in these schools are not native speakers of respective language.

The participation rate in grade 9 in Finnish and Swedish speaking schools respectively at each year of investigation was, 1994: 88.7/92.1 %, 1998: 88.9/94.8 %, 2006: 84.5/86.5 %, 2014: 85.5/78.5 %.

The HBSC questionnaire covers subjective health, including mental health complaints. The questions have been translated to Finnish and Swedish from the prototype questionnaire in English using a standard protocol, including back translation to English at the final step.

In the international reports on the HBSC-study only data from the Finnish speaking schools are included, and eight items are used to measure psychosomatic problems. While the Finnish data include the item depressed as an indicator of emotional problems, the item feeling low is used in other participating countries. Albeit not included in the international data set, Finnish data on feeling low are available in addition to the data on feeling depressed. This distinctive feature of the Finnish data set is utilized in the present paper by including both the item depressed and the item feeling low in the analyses. Hence, for the specific purpose of this study, nine items on psychosomatic problems are used in the psychometric analyses.

In Table 7 (appendix) the nine questions on psychosomatic problems are presented in English, Finnish, and Finland Swedish. For reasons of comparability of phrasing, the questions are also presented in the Swedish language spoken in Sweden since the Finland Swedish and Swedish languages differ somewhat concerning idiomatic expressions.

The response categories for all of the nine items, which are in the form of questions, are 'About every day', 'More than once a week', 'About once a week', 'About once a month' and 'Seldom or never'. The categories are ordered in terms of implied frequency and the higher frequency, the more the psychosomatic problems.

The phrasing of the questions has been the same at all years of investigations, except for the item feeling low in the Finland Swedish questionnaire which was changed in 2014.

2.1 The Model of the Analysis

The items were analysed using the polytomous Rasch model (Andrich 1978; Hagquist et al. 2009). The model was used to examine to what extent the data met measurement requirements of invariance and proper item categorisation enabling invariant comparisons along the latent trait as well as across time. Since invariance is an integral property of the Rasch model, misfit between the data and the model indicates lack of invariance in the functioning of a measurement instrument (Hagquist et al. 2009).

In the Rasch analysis the focus is on the operating characteristics of the items along the whole continuum of a latent trait, not on a single summary measure. In the polytomous Rasch model that comprises more than two ordered categories there are two item parameters, the item location and the threshold parameters. The thresholds are partitioning the latent continuum of an item into ordered categories, one more than the number of thresholds (Andrich 1978). The item parameters are independent of the person distribution because they are estimated conditioning on the sum of the raw scores, which is a sufficient statistic (Andrich 2011). Inversely, the person parameters do not depend on which items are used for the estimation.

It has to be emphasized that the Rasch model provides invariant measures only if the data fit the model. Notably, the view of the relationship between the data and the model is opposite to traditional statistical analysis in which variables are added in order to describe the data (Andrich 2013).

The Expected Value Curve (EVC) is an excellent graphical tool to check the fit of the data to the Rasch model, complementing formal test statistics, such as the chi-square-statistic (Hagquist and Andrich 2015). In a polytomous item the slope of the EVC is a function of the distances between the item thresholds, i.e. the closer the thresholds are located, the steeper becomes the slope.

The EVC predicts the responses to the items as a function of the items and persons locations on the latent trait. These expected values are compared with the observed values. Ideally, the observed values should fit perfectly with the EVC.

Lack of invariance across different sample groups, e.g. gender and age cohorts, is commonly called Differential Item Functioning (DIF). This means that members of one sample group score higher on an item than members of another group, given the same location on the latent trait. This difference implies that two different EVCs are required, e.g. one for boys and one for girls, to predict the responses on an item. If the difference is the same along the latent trait implying parallel EVCs, the DIF is referred to as uniform; if the difference varies along the latent trait implying non-parallel EVCs, it is referred to as non-uniform (Andrich and Hagquist 2015).

While the EVC-procedure for detecting DIF is fairly new, there are also other approaches which estimate and compare item parameters from different groups, which are complementary to the graphical displays. Item location values for two or more groups, e.g. boys and girls, may be compared to measure the magnitude of uniform DIF for a specific item, while comparisons of the slope values between two groups provide a measure of the magnitude of non-uniform DIF (Andrich and Hagquist 2012).

In DIF-analysis of an item set, several items may show evidence of DIF, comprising real DIF items as well as artificial DIF items. Real DIF is inherent to an item and affects the person measures, while artificial DIF does not since it is an artefact of the procedure

for identifying DIF. The concepts of real and artificial DIF have recently been introduced and described using dichotomous (Andrich and Hagquist 2012) as well as polytomous data (Andrich and Hagquist 2015) and focusing both uniform and non-uniform DIF (Hagquist and Andrich 2015).

In the present study graphical representations of DIF with reference to the EVC are complemented with two-way analysis of variance of standardised residuals (Hagquist and Andrich 2004). In an analysis of potential gender DIF that procedure examines each item with respect to a class interval main effect, a gender effect and a class interval by gender interaction effect.

The Rasch analysis was performed with the software RUMM2030 (Andrich et al. 2014).

RUMM2030 uses a pairwise conditional method of estimation for the item parameters in which person parameters are eliminated and the person estimates were obtained using a weighted likelihood method which reduces the bias in the person estimates taking the item parameters as known (Andrich et al. 2014). In the present study the unrestricted version of the polytomous Rasch model was applied, referred to as the partial credit model (Masters 1982). In this version the distances between thresholds may vary across items, because the model allows the thresholds to take different values across the items.

2.2 Analysis

The summated raw scores of the students' responses across the nine items were non-linearly transformed to a linear variable using Rasch analysis. Lower person estimate values on this psychosomatic problems variable mean higher degree of psychosomatic problems.

A test for multidimensionality was conducted based on residual principal component analysis providing two subsets of items. Independent t-tests of differences between person locations values from these two subsets of items were carried out.

Response dependency was examined by analysing the correlations between the item residuals.

In addition to a general level Rasch analysis among grade 9 students, a finer level analysis was conducted consisting of two parts:

- a) Analysis of DIF across years of investigations among four subgroups of students, differing with respect to gender and language spoken at school.
- b) Analysis of differences in item location values between the two content-related items, depressed and feeling low, among four subgroups of students, differing with respect to gender and language spoken at school.

The DIF-analyses were conducted at the item-level as well as at the person level. To provide measures of the magnitude of DIF, the items showing real DIF were resolved, which means that the DIF item was split into multiple group specific items, e.g. one for each year of investigation, in which the values for the excluded groups were treated as structural missing. After that resolution of the DIF item, the item parameters were compared between the groups. If there are several items showing DIF in the original analysis, the DIF items should not be resolved simultaneously but sequentially

starting with the item showing the greatest F-value. The magnitude of DIF was also examined at the person level, e.g. by comparing the group differences in person mean values before and after resolution of the DIF-item.

Resolving DIF-items may be a way of taking account of DIF, but another option could be to simply remove the real DIF item(s). While removal of an item may decrease the reliability and person separation, resolving a DIF item will have only a small, if any, impact on the Person Separation Index (PSI). Therefore, often it may be preferable to resolve an item instead of removing it. However, resolving an item, like removing an item, may affect the validity implying that this procedure is only justified if the source of DIF can be shown to arise from some source irrelevant to the variable of assessment and therefore deemed dispensable.

In the present study, changes in mean person values were compared for three different item sets: the original set of nine items, an item set with the worse DIF-item resolved and an item set in which the worse DIF-item was removed.

The DIF-analyses were conducted with a sample size adjusted to the value of the order of 1000 with the Bonferroni adjustment (Bland and Altman 1995) of significance values applied for a Type I error level of 0.05.

At the final stage a set of eight items without the item depressed was analysed with respect to item fit, threshold ordering, DIF and local independency.

3 Results

3.1 Grade 9 Students – Original Item Set with Nine Items

The analysis of the entire sample of grade 9 students shows a high value of the PSI (0.79798), two items with significant misfit between observed and expected values and no item with disordered thresholds. There were also indications of local dependency: T-tests based on two subsets of items grouped by principal component analysis of residuals indicated that multidimensionality could not be ruled out. Also, analysis of residual correlations indicated response dependency: there was a correlation of the order >0.3 between items depressed and feeling low which means that the response on one item may have affected the response on the other item.

Analysis of DIF shows lack of invariance, in particular for the item depressed across years of investigations (uniform and non-uniform). There is also some gender DIF, but very little DIF across language, i.e. the items worked in a similar way for grade 9 students in Finnish and Swedish speaking schools.

In the following, the results from the finer level analysis are reported using graphical representations of DIF. The displays of DIF-items are presented along with graphs in which the DIF items are resolved, providing estimates of the magnitude of DIF based on the differences between item parameters for different groups.

In Fig. 1a–d Expected Value Curves for the item depressed in Finnish speaking schools are shown, distributed by years of investigations and gender.

Figure 1a–d shows that the expected value for the item depressed is higher in the years 2006 and 2014 than in 1994 and 1998 among both boys and girls in grade 9 in Finnish speaking schools. This means that regardless of the students' locations on the latent variable, at the two first years of investigations the students score lower values (=more

Fig. 1 Differential Item Functioning across years of investigations for item Depressed before (a, c) and after DIF is resolved (b, d). HBSC data from Finland 1994 – 2014. Grade 9 students in Finnish speaking schools distributed by gender

frequent problems) than at the two last years. This difference in functioning across years is confirmed when DIF is resolved. The estimates of the item parameters differ between the first two and the last two years of investigations. This applies to the location values as well as the slope values which are indicative of uniform as well as non-uniform DIF.

In Fig. 2a–d the relative frequencies of responses on the items depressed and feeling low in Finnish speaking schools are shown, distributed by years of investigations and gender.

The relative frequencies for the item depressed shown in Fig. 2a–d reflect the DIF-analyses presented in Fig. 1a–d. Among both genders, the item depressed and the item feeling low show opposite patterns with respect to changes over time. The proportion of students feeling depressed regularly is decreasing across years of investigations while the proportion feeling depressed seldom or never is increasing. In contrast, the proportion of students feeling low regularly is increasing across years of investigation while the proportion feeling low seldom or never is decreasing.

In Fig. 3a–d DIF for the item depressed in Swedish speaking schools is shown, distributed by years of investigations and gender.

Figure 3a–d shows that the DIF across years of investigations is less pronounced among boys and girls in grade 9 in Swedish speaking schools compared to Finnish speaking schools. Comparing the first and last years of investigations, in 1994 boys and girls tend to score lower values (=more frequent problems) than 2014 on the item depressed, regardless of the students' locations on the latent variable. This difference in functioning across years is confirmed by resolving the item depressed. The estimates of the item parameters differ between the first and the last years of investigation. This applies to the locations values as well as the slope values which are indicating uniform as well as non-uniform DIF.

In Fig. 4a–d the relative frequencies of responses on the items depressed and feeling low in Swedish speaking schools are shown, distributed by years of investigations and gender.

Fig. 2 The relative frequencies for item depressed and item feeling low. HBSC data from Finland 1994 – 2014 in Finnish speaking schools. Grade 9 students distributed by years of investigations and gender

The relative frequencies shown in Fig. 4a–d reflect the DIF-analyses presented in Fig. 3a–d. Among girls the item depressed and the item feeling low show slightly opposite patterns with respect to changes over time. The proportion of girls feeling

Fig. 3 Differential item functioning across years of investigations for item depressed before (a, c) and after DIF is resolved (b, d). HBSC data from Finland 1994 – 2014. Grade 9 students in Swedish speaking schools distributed by gender

Fig. 4 The relative frequencies for item Depressed and item Feeling low. HBSC data from Finland 1994 – 2014 in Swedish speaking schools. Grade 9 students distributed by years of investigations and gender

depressed seldom or never is increasing, while the corresponding proportion for feeling low is decreasing. Among boys the patterns are less clear.

In Table 1 item location values for the original set of nine items are shown for the entire sample of grade 9 students and distributed by gender and by type of school (language).

Table 1 Item location values for nine items on psychosomatic problems

Item label	All	Boys	Girls	Finnish speaking	Swedish speaking
Headache	0.230	0.072	0.294	0.256	0.156
Stomachache	-0.349	-0.488	-0.199	-0.365	-0.302
Backache	-0.098	0.017	-0.197	-0.113	-0.050
Depressed	-0.274	-0.331	-0.242	-0.275	-0.270
Irritable/Bad temper	0.624	0.632	0.661	0.608	0.675
Nervous	0.228	0.363	0.129	0.257	0.134
Sleep difficulties	0.079	0.273	-0.063	0.076	0.093
Dizzy	-0.220	-0.171	-0.254	-0.191	-0.314
Feeling low	-0.220	-0.368	-0.129	-0.252	-0.122

HBSC data from Finland 1994 – 2014 for all grade 9 students, and distributed by gender and by language of the school respectively. The lower the value of an item, the more severe health problems

Comparing the items depressed and feeling low, Table 1 shows that the internal relationship between the two items differs across subgroups. Among grade 9 students as a whole, the location value of depressed is lower than the corresponding value of feeling low. Given the ordering of the latent scale values this means that depressed is a more severe item indicating more severe problems, while feeling low is indicating less severe problems. This pattern holds also for three of the four subgroups. A reversed ordering is shown for boys in grade 9. Regardless of their internal relationship, both the item depressed and the item feeling low belong to the group of items representing more severe problems within the set of nine items.

In Table 2 item location values for the original set of nine items are shown for different subgroups of grade 9 students.

Comparing the items depressed and feeling low, Table 2 shows a complex pattern. There is a reversed gender pattern for grade 9 students in Finnish speaking schools. Among girls the item depressed is more severe than the item feeling low, while among boys the relationship is the opposite. For Swedish speaking schools the gender pattern is consistent. Among boys as well as girls the item depressed is more severe than the item feeling low.

3.2 Grade 9 Students – Revised Item Set with Eight Items

The value for the PSI is 0.77532 in the revised item set without item depressed. The analysis of the entire sample shows no disordered thresholds. In this set there is only small, if any, evidence of response dependency, indicated by all item residual correlations being below the value of 0.3. Based on the outcomes from principal component analyses of the item residuals two sets of items were compared, and the differences between person locations values were assessed with t-tests. Comparisons between set 1

Table 2 Item locations for nine items on psychosomatic problems

Item label	Finnish speaking		Swedish speaking	
	Boys	Girls	Boys	Girls
Headache	0.094	0.323	0.020	0.208
Stomachache	-0.523	-0.202	-0.398	-0.185
Backache	0.026	-0.220	-0.008	-0.128
Depressed	-0.334	-0.243	-0.318	-0.237
Irritable/Bad temper	0.606	0.654	0.707	0.690
Nervous	0.393	0.160	0.287	0.015
Sleep difficulties	0.290	-0.078	0.235	-0.014
Dizzy	-0.132	-0.232	-0.291	-0.331
Feeling low	-0.421	-0.162	-0.234	-0.018

HBSC data from Finland 1994 – 2014. Grade 9 students distributed by gender and language of the school. The lower the value of an item, the more severe health problem

(Irritable, Nervous, Sleep difficulties, Feeling low) and set 2 (Headache, Stomach ache, Backache, Dizzy) indicated no, or only minor, evidence of multidimensionality (at 0.05: 5.83 %; at 0.01: 0.81 %).

In Fig. 5 the person-item threshold distribution for grade 9 of the revised eight item set without the item feeling depressed is shown.

Figure 5 shows that the person distribution is skewed to the right with a person mean value of about 0.9. Most of the item threshold values are located at the lower part of the variable (=more psychosomatic problems), while there are only a few threshold values at the upper part of the variable (=less psychosomatic problems).

In Table 3 item location and threshold values from the analyses for grade 9 of the revised set of eight items without the item depressed are shown.

Table 3 shows that almost all items fit poorly according to the p -values based on the unadjusted sample size, but that the items fit much better based on the adjusted sample size of the order of 1000. No item shows disordered item thresholds which confirms that the ordering of the response categories is working as intended.

In Table 4 analyses for grade 9 of DIF of the revised set of eight items without the item feeling depressed are shown.

Table 4 shows that there is no item showing DIF by year or by language in the analyses based on the adjusted sample size. There are some items showing gender DIF. The worse DIF items are stomach ache and feeling low, both with higher expected values (=lower frequency) for boys than girls given the same location on the latent variable.

3.3 Grade 9 Students – Comparisons of Person Measures Between Different Item Sets

In Table 5 the person mean values of psychosomatic problems are shown for different groups of grade 9 students along with PSI values, before, after resolving DIF and after removing the item depressed.

Fig. 5 Person-Item threshold distribution, grade 9, revised set of eight items

Table 3 HBSC data from Finland 1994 – 2014. Grade 9

Item label	Location estimate	P-val Chi2-test		Thresholds			
		<i>n</i> = 8254	<i>n</i> = 1000	1	2	3	4
Headache	0.193	0.183338	0.996979	−1.14558	−0.51588	0.158633	1.502825
Stomachache	−0.382	0.000000	0.670358	−1.25995	−0.40544	−0.339010	2.004403
Backache	−0.134	0.000000	0.127965	−0.53300	−0.19573	−0.08344	0.812166
Irritable/Bad temper	0.579	0.000000	0.051205	−1.8163	−0.68012	0.435084	2.061342
Nervous	0.191	0.000000	0.032947	−1.50238	−0.50010	0.407557	1.594920
Sleep difficulties	0.043	0.000038	0.885315	−0.76469	−0.16973	0.358048	0.576369
Dizzy	−0.253	0.000058	0.894271	−0.64832	−0.03213	0.229403	0.451042
Feeling low	−0.235	0.000000	0.020184	−0.99383	−0.11692	0.058173	1.052566

Estimates of individual item parameters in the revised set of eight items. Chi²-test based on original sample size and adjusted sample size respectively

Comparing changes in person mean values between the years 1994 and 2014 for the original item set of nine items, Table 5 indicates lower mean values over time for girls, in particular in Finnish speaking schools, i.e. increasing psychosomatic problems. Resolving the DIF-item depressed for year amplifies the differences in mean values for girls between the first and last years of investigations, implying increasing psychosomatic problems. Among boys the changes are minor. Before resolving the DIF-item the mean values for Finnish speaking boys are the same for 1994 and 2014, and for Swedish speaking boys marginally higher (=less problems) 2014 than 1994. After resolving the DIF-item there is a small difference for Finnish speaking boys, with lower values (=more problems) in 2014 than 1994. Removing the item depressed instead of resolving the item has roughly the same impact on person measurement. While removal of the item depressed decreased the PSI, the values stayed about the same for the item set with the item depressed resolved for DIF as for the original set of nine items.

3.4 Grade 5, 7 and 9 Students – Revised Item Set with Eight Items

In Table 6 item analyses for grades 5, 7 and 9 of DIF of the revised set of eight items without the item feeling depressed are shown.

Table 6 shows that the item feeling low has significant class interval effects in all DIF analyses conducted. This item is over discriminating meaning that persons at the lower end of the variable are scoring lower (=higher frequency) than expected while persons at the upper end of the variable are scoring higher (=lower frequency) than expected. The item stomach ache shows gender DIF with higher expected values for boys than girls. The items backache and dizzy show DIF with respect to grade with higher expected values for grade 5 than grade 7, and for grade 7 compared to grade 9.

Table 4 HBSC data from Finland 1994 – 2014, Grade 9

Item label	Division by year of investigation			Division by gender			Division by class interval			Division by language		
	Year			Gender			Class interval			Class interval		
	Class interval	Year	Year by class interval	Class interval	Gender	Gender by class interval	Class interval	Gender	Gender by class interval	Class interval	Language	Language by class interval
Headache	0.999611	0.981360	1.000000	0.999562	0.006570	0.816186	0.999582	0.373619	0.999869	0.999582	0.373619	0.999869
Stomachache	0.775246	0.689176	1.000000	0.744219	0.000002	0.919534	0.765839	0.956550	0.994553	0.765839	0.956550	0.994553
Backache	0.166702	0.017380	0.999982	0.157695	0.000078	**N/Sig	0.164211	0.697419	0.999949	0.164211	0.697419	0.999949
Irritable/Bad temper	0.011735	0.036193	0.999999	0.011066	0.210731	**N/Sig	0.010885	0.335257	0.984087	0.010885	0.335257	0.984087
Nervous	0.006060	0.473855	1.000000	0.004907	0.033598	0.728404	0.005150	0.210639	0.964061	0.005150	0.210639	0.964061
Sleep difficulties	0.947487	0.496008	1.000000	0.942543	0.000171	**N/Sig	0.944788	0.964500	0.999949	0.944788	0.964500	0.999949
Dizzy	0.994045	0.134411	0.999998	0.99379	0.618213	0.997543	0.993756	0.182953	0.998220	0.993756	0.182953	0.998220
Feeling low	0.005868	0.661600	0.999999	0.004514	0.000000	**N/Sig	0.004955	0.039680	0.998121	0.004955	0.039680	0.998121

Analysis of variance of residuals for test of DIF across years of investigations, gender and language as well as tests of class interval fit. Revised set of eight items. Number of class intervals = 10. Adjusted sample size for test of fit ($n = 1000$); Bonferroni adjusted significance level: 0.002083. Items showing significant DIF marked in bold

Table 5 Mean of person estimate values across years of investigations and person separation index for composite measure of psychosomatic problems

Item label	Finnish speaking Boys				Finnish speaking Girls				Swedish speaking Boys				Swedish speaking Girls			
	Original 9 items	Item Depressed resolved	Item Depressed removed	Original 9 items	Item Depressed resolved	Item Depressed removed	Original 9 items	Item Depressed resolved	Original 9 items	Item Depressed resolved	Item Depressed removed	Original 9 items	Item Depressed resolved	Original 9 items	Item Depressed resolved	Item Depressed removed
Person Separation Index (PSI)	0.77729	0.77867	0.75737	0.80862	0.81225	0.78485	0.76249	0.76245	0.73836	0.78935	0.79029	0.75417				
Mean 1994	1.305	1.462	1.305	0.842	0.961	0.858	1.153	1.264	1.125	0.713	0.786	0.675				
Mean 1998	1.315	1.425	1.275	0.772	0.847	0.755	1.371	1.452	1.301	0.830	0.898	0.783				
Mean 2006	1.196	1.249	1.106	0.652	0.667	0.581	1.174	1.243	1.098	0.625	0.682	0.575				
Mean 2014	1.306	1.367	1.224	0.608	0.623	0.540	1.204	1.271	1.136	0.569	0.612	0.509				
Difference 94–14	−0.001	0.095	0.081	0.234	0.338	0.318	−0.051	−0.007	−0.011	0.144	0.174	0.166				

HBSC data from Finland 1994–2014. Grade 9 students distributed by gender and language of the school. The lower the value, the higher degree of psychosomatic problems

Table 6 HBSC data from Finland 1994 – 2014, Grades 5, 7 and 9

Item label	Division by year of investigation				Division by gender			Division by language			Division by grade		
	Class interval	Year	Year by class interval	Class interval	Gender	Gender by class interval	Class interval	Language	Language by class interval	Class interval	Grade	Grade by class interval	Grade
Headache	0.999890	0.932803	1.000000	0.999879	0.108411	0.994678	0.999880	0.155280	0.999666	0.999886	0.932053	1.000000	
Stomachache	0.938656	0.744625	1.000000	0.932219	0.000146	**N/Sig	0.935123	0.899010	0.999962	0.936471	0.526673	0.999989	
Backache	0.696122	0.015013	1.000000	0.686776	0.004942	**N/Sig	0.690872	0.629786	0.999952	0.686103	0.000756	0.999949	
Irritable/Bad temper	0.112544	0.027356	1.000000	0.107822	0.129392	**N/Sig	0.108307	0.780521	0.999959	0.109802	0.106835	0.999972	
Nervous	0.023883	0.688408	1.000000	0.020342	0.090687	0.950851	0.020175	0.081658	0.895159	0.021965	0.443897	0.999952	
Sleep difficulties	0.817413	0.491600	1.000000	0.807174	0.018678	**N/Sig	0.808939	0.602895	0.999961	0.810693	0.020646	**N/Sig	
Dizzy	0.995559	0.797825	1.000000	0.995245	0.914168	0.999929	0.995220	0.269483	0.999822	0.995188	0.001764	1.000000	
Feeling low	0.002004	0.172085	1.000000	0.001690	0.005193	**N/Sig	0.001580	0.002110	0.999889	0.001858	0.020525	**N/Sig	

Analysis of variance of residuals for test of DIF across years of investigations, gender and language as well as tests of class interval fit. Revised set of eight items. Number of class intervals = 10. Adjusted sample size for test of fit ($n = 1000$). Bonferroni adjusted significance level: 0.002083. Items showing significant DIF marked in bold

4 Discussion

The purpose of the present paper was to analyse the psychometric properties of a composite measure of psychosomatic problems based on data from Finland, particularly with respect to Differential Item Functioning (DIF) across years of investigations.

The results from the current study clearly indicate that a nine items measure of psychosomatic problems does not work invariantly across years of investigations. In particular, the item depressed shows DIF across years of investigations. This item works quite differently at the first year of investigation compared to the last year showing higher expected values 2014 (=less frequent problems) than 1994, given the same location of the persons on the latent trait.

The DIF-pattern also holds for all subgroups but is most pronounced for students in Finnish speaking schools, in particular girls. This DIF does affect the measure of change in psychosomatic problems between 1994 and 2014. Resolving the item depressed for year DIF changes (increases) the difference in person mean values between the two years, which also confirms that the DIF is real, not artificial (Andrich and Hagquist 2015). Hence, the DIF suppresses an increasing trend of psychosomatic problems among adolescents.

The analysis of the internal relationship between the items depressed and feeling low also shows some unexpected results. Usually, feeling depressed is assumed to reflect more severe health problems than feeling low. However, the results reported in this paper partly contradict that hypothesis. Among boys in grade 9 the item feeling low turns out to be a more severe item than the item feeling depressed.

The direction of the DIF across years of investigations for the item depressed may come as a surprise. At first glance it would be reasonable to expect an opposite direction of DIF across years because of greater proneness to report emotional problems and decreasing stigmatization over time. Given the increasing focus on adolescent emotional problems, rising awareness of the actual meaning of being depressed could be an alternative explanation of the decreasing number of students feeling depressed.

Since the DIF is shown to affect the trend results, consideration of different options to address the described problems is advisable. Some knowledge about the sources of the DIF is desirable in order to guide a decision about what to do. From a content perspective the key issue is if, and which, societal changes in attitudes, values and knowledges may explain the time DIF for the item depressed.

As a methodological rule of thumb, neither removing nor resolving items is justified solely based on statistical criteria e.g. significant DIF. Evidence of the sources of DIF, e.g. whether the perceived meaning of the item depressed has changed over the years among Finnish adolescents, would facilitate a decision about what to do.

Since no external information allowing for firm conclusions is available about the sources of the detected DIF, in the following a possible hypothesis will be discussed with reference to changes in the Finnish school curriculum. In 2001 health education became an independent school subject in both

compulsory schools and in upper secondary schools in Finland. Although health issues had been taught in Finnish schools for many decades, mainly in connection with physical education, as an independent school subject health education was reoriented towards a more comprehensive understanding of health (Paakkari and Välimaa 2013). Because of an increasing number of health education courses, there are nowadays more possibilities to focus topics on mental health (Finnish National Board of Education 2014). These changes as well as an increasing focus on mental health issues in media may have raised the awareness among adolescents about the actual meaning of different mental health concepts and helped to distinguish between them.

In general, resolving instead of removing DIF-items is preferable because removal of an item is likely to decrease the precision of measurement whilst resolving an item is not. Recalling that the Finnish set of eight psychosomatic items that have been used in the international HBSC data set deviates from the HBSC protocol, removing the item depressed seems to be the best option in the present analyses of the Finnish nine items set. That action would also provide a set of eight items in better accordance with the content of the measure of psychosomatic problems described in the international HBSC protocol. Furthermore, as a whole the revised set of eight items without the item depressed works well among grade 9 students, but there are some items showing evidence of DIF across genders. Whether these DIF items should be resolved or retained in their original format is a question to be addressed in further analyses.

4.1 Conclusion

According to the result of the present study a composite measure of nine psychosomatic problems including item depressed does not work invariantly across years of investigations among grade 9 students in Finland. In what ways the results and conclusions presented in the current paper should affect the use of the Finnish data as part of the international HBSC data set is an issue to be addressed by the HBSC data management group among others. In doing that, finer level psychometric analyses also including grades 5 and 7, need to be taken into consideration. The study challenges the traditional view of interpreting results from trend studies dealing with mental health of adolescents solely from the perspective of whether the questions have been retained in their original format and/or whether data were collected by standardized procedures.

Acknowledgments The research was funded by Forte: the Swedish Research Council for Health, Working Life and Welfare (Program grant number: 2012–1736).

A previous version of this paper was presented at the 7th Excellence in Pediatrics Conference, 10–12 December 2015, London UK.

Appendix

Table 7 Items on psychosomatic problems

English prototype	Finnish (back)translation	Finnish	Finland Swedish	Swedish
M97, Headache	headache	päänsärky	huvudvärk	huvudvärk
M98, Stomachache	stomachache	vatsakipu	ont i magen	ont i magen
M99, Backache	backache	selkäkipu	ont i ryggen	ont i ryggen
M1000, Feeling low	feeling depressed	masentuneisuus	känner mig deprimerad	känt mig nere
M101, Irritability or bad temper	irritability or bad tempered	ärtäisyys tai pahantulisuus	irriterad eller på dåligt humör	varit irriterad eller på dåligt humör
M102, Feeling nervous	feeling nervous	hermostuneisuus	känner mig nervös	känt mig nervös
M103, Difficulties in getting to sleep	difficulties in getting to sleep	vaikeuksia päästä uneen	har svårt att somna	haft svårt att somna
M104, Feeling dizzy	feeling dizzy	huimauksen tunne	känner mig yr	känt mig yr
FIN53, [EXTRA] alakuloisuus (feeling low)	feeling low	alakuloisuus	känner mig deppig 2014: känner mig nedstämd	

Eight HBSC-items and one extra item with five response categories included in the Finnish survey

References

- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4), 561–573.
- Andrich, D. (2011). Rating scales and Rasch measurement. *Expert Review of Pharmacoeconomics & Outcomes Research*, 11(5), 571–585.
- Andrich, D. (2013). The Legacies of R. A. Fisher and K. Pearson in the application of the polytomous rasch model for assessing the empirical ordering of categories. *Educational and Psychological Measurement*, 73(4), 553–580.
- Andrich, D., & Hagquist, C. (2012). Real and artificial differential item functioning. *Journal of Educational and Behavioral Statistics*, 37(3), 387–416.
- Andrich, D., & Hagquist, C. (2015). Real and artificial differential item functioning in polytomous items. *Educational and Psychological Measurement*, 75(2), 185–207. doi:10.1177/0013164414534258
- Andrich, D., Sheridan, B., & Luo, G. (2014). *RUMM2030: A windows interactive program for analysing data with Rasch unidimensional models for measurement*. Perth: RUMM Laboratory.
- Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: the Bonferroni method. *British Medical Journal*, 310, 170.
- Bor, W., Dean, A. J., Najman, J., & Hayatbakhsh, R. (2014). Are child and adolescent mental health problems increasing in the 21st century? A systematic review. *Australian and New Zealand Journal of Psychiatry*, 48(7), 606–616. doi:10.1177/0004867414533834.
- Cieza, A., Oberhauser, C., Bickenbach, J., Jones, R. N., Üstün-Bedirhan, T., Kostanjsek, N., et al. (2015). The English are healthier than the Americans: really? *International Journal of Epidemiology*, 44(1), 229–238. doi:10.1093/ije/dyu182.
- Currie, C. et al. (Eds) (2012). Social determinants of health and well-being among young people. Health Behaviour in School-aged Children (HBSC) study: International report from the 2009/2010 survey. Copenhagen: WHO Regional Office for Europe. (Health Policy for Children and Adolescents, No. 6).
- Finnish National Board of Education. (2014). *National core curriculum for basic education 2014*. Helsinki: Finnish National Board of Education.
- Hagquist, C., & Andrich, D. (2004). Is the sense of coherence-instrument applicable on adolescents? A latent trait analysis using Rasch-modelling. *Personality and Individual Differences*, 36(4), 955–68. doi:10.1016/S0191-8869(03)00164-8
- Hagquist, C., & Andrich, A. (2015). Determinants of artificial DIF – a study based on simulated polytomous data. *Psychological Test and Assessment Modeling*, 57(3), 342–376.
- Hagquist, C., Bruce, M., & Gustavsson, J. P. (2009). Using the rasch model in nursing research: An introduction and illustrative example. *International Journal of Nursing Studies*, 46(3), 380–93. doi:10.1016/j.ijnurstu.2008.10.007
- Kreiner, S., & Christensen, K. B. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika*, 79(2), 210–231. doi:10.1007/s11336-013-9347-z.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174.
- Ottová-Jordan, V., Smith, O. R., Gobina, I., Mazur, J., Augustine, L., Cavallo, F., et al. (2015). Trends in multiple recurrent health complaints in 15-year-olds in 35 countries in Europe, North America and Israel from 1994 to 2010. *European Journal of Public Health*, 25(2), 24–27. doi:10.1093/eurpub/ckv015.
- Paakkari, L., & Välimaa, R. (2013). Ethical issues in the teaching and learning of health topics in schools: the conceptions of teacher trainees. *Teaching and Teacher Education*, 34, 66–76. doi:10.1016/j.tate.2013.04.004.
- Ravens-Sieberer, U., Erhart, M., Torsheim, T., Hetland, J., Freeman, J., Danielson, M., & Thomas, C. (2008). An international scoring system for self-reported health complaints in adolescents. *European Journal of Public Health*, 18(3), 294–299. doi:10.1093/eurpub/ckn001.
- Statistics Finland (2014). *Preliminary population statistics*. Helsinki: Official Statistics of Finland, 2014, March. http://tilastokeskus.fi/til/vamuu/2014/03/vamuu_2014_03_2014-04-24_en.pdf. Accessed 21 Oct 2015.