

Differential Item Functioning of a Family Affluence Scale: Validation Study on Data from HBSC 2001/02

C. W. Schnohr · S. Kreiner · E. P. Due · C. Currie · W. Boyce · F. Diderichsen

Accepted: 10 November 2007
© Springer Science+Business Media B.V. 2007

Abstract Methodology for making cross-national comparisons is an area of increasing interest in social and public health related research. When studying socio-economic differences in health outcomes cross-nationally, there are several methodological issues of concern, especially when data is derived from self-reported questionnaires. Health Behaviour in School-aged Children (Currie et al. 1998) is a WHO cross-national study using school samples. HBSC provides comparable data, and thereby a unique opportunity to study associations between social indicators and health outcomes within an international perspective. In 2001/02 data was collected from a total of 162,323 children in 32 countries (Austria, Belgium, Canada, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Greenland, Hungary, Ireland, Israel, Italy, Latvia, Lithuania, Macedonia, Malta, Netherlands, Norway, Poland, Portugal, Russia, Slovenia, Spain, Sweden, Switzerland, Ukraine, United Kingdom, USA.). Studies of social inequalities requires that a comparable measure of socio-economic position (SEP) is in use. HBSC has developed a proxy for social position measuring material wealth, the Family Affluence Scale (FAS). This paper studies FAS and whether it is comparable across population subgroups defined by country, age and gender. Initial analysis revealed that an item measuring perceived family wealth was a valid FAS item. Including this item in the FAS score will improve the reliability of FAS. Graphical log-linear Rasch models (GLLRM) showed that FAS contain differential item functioning (DIF) with respect to country, age, and gender as well as local

C. W. Schnohr (✉) · E. P. Due · F. Diderichsen
Department of Social Medicine, Institute of Public Health, University of Copenhagen,
Oester Farimagsgade 5, P.O. Box 2099, Copenhagen 1014, Denmark
e-mail: c.schnohr@pubhealth.ku.dk

S. Kreiner
Department of Biostatistics, Institute of Public Health, University of Copenhagen, Copenhagen,
Denmark

C. Currie
Child and Adolescent Health Research Unit, University of Edinburgh, Edinburgh, Scotland

W. Boyce
McArthur Hall, Queen's University, Kingston, Ontario, Canada

dependency (LD) between items. During the analysis, test equating techniques were used to adjust for the test bias generated by DIF. We recommend that the equated scores are used whenever FAS is included as a variable. This study suggests that HBSC-FAS should contain five items (additional item: perceived family wealth) when analysing data from HBSC 2001/02, and furthermore that each country should adjust for the DIF or make use of the converted FAS scores provided. If using FAS as a proxy for social position at an international level, it is not advised to compare the absolute levels of FAS, but weigh the scale by ridit transformation.

Keywords Family affluence scale · Differential item functioning · Health Behaviour in School-aged Children (HBSC) · Comparability · Validity

Abbreviations

HBSC	Health Behaviour in School-aged Children
SEP	Socio Economic Position
FAS	Family Affluence Scale
HASC	Home Affluence Scale
DIF	Differential Item Functioning
LD	Local Dependence
GLLRM	Graphical Log-Linear Rasch Models
CLR	Conditional Likelihood Ratio test
PFW	Perceived Family Wealth
IRT	Item Response Theory

1 Introduction

Social inequality in health is a property of all societies where it has been examined. In the search for effective ways of reducing them, there is a growing interest in analysing the international variations in the size of those inequalities (Mackenbach and Bakker 2002). Policies and interventions with a potential to reduce inequalities in health can seldom be evaluated by randomized controlled trials. We therefore have to rely on the natural experiments created by variations in both policy and the size of inequalities across societies (Whitehead et al. 2000). Most comparative studies have so far been studying health inequalities in adult populations, and studies on child and adolescents will therefore be informative.

Specific problems pertain to measuring social position of children by asking children or adolescents about their parents' social position (Currie et al. 1997). Questionnaires to adolescents on their parents' education, occupation or income usually result in low completion rates and high misclassification rate (Currie et al. 1997; Lien et al. 2001; Wardle et al. 2002), so various alternatives have been examined. In the HBSC study, a scale used to measure family affluence, the Family Affluence Scale (FAS) has earlier included items on telephone ownership, car ownership and numbers of spent holidays etc. The Home Affluence Scale (HASC) (Wardle et al. 2002) has used housing tenure, computer ownership and school meals among others. These items can be used either as separate

independent (albeit correlated) exposures or they can be assumed to be items of some underlying (latent) construct such as material condition of the household.

Adding items together, as is often done (Currie et al. 1997; Torsheim et al. 2004), implies an assumption of cumulative effects of the different items, and that items can be combined to a reliable scale measuring an underlying construct. Cross-national use of scales constructed by several items potentially suffer from differential item functioning (DIF) (Batista-Foguet et al. 2006; Björner et al. 1998b), and from the perspective of comparative studies the question arises whether such a scale will be comparable across population subgroups defined by country, age and gender. This paper studies the FAS in the large cross-national study Health Behaviour in School-aged Children (HBSC) (Currie et al 1998).

1.1 The Health Behaviour in School-aged Children and the Family Affluence Scale

HBSC is an international study including adolescents from 32 countries in Europe, Israel, Canada and the US. The study provides comparable data on young people's health and lifestyle from countries with different societal and political systems.

The HBSC study consists of repeated cross-sectional surveys among 11-, 13- and 15-year-old school children in representative samples of schools in the participating countries. The students answer a standardized questionnaire during a school lesson after instruction from the teacher. HBSC has been collecting data on adolescents every fourth year since 1982, the most recent survey in 2006 included 40 countries and regions. The magnitude and the general comparability of the study make it an important task to enhance the possibilities of carrying out comparative studies on the data.

In the HBSC-study SEP is measured by a scale on family wealth; the Family Affluence Scale (FAS). Assessing adolescent's absolute socio-economic status based on material markers provides an alternative from the more traditional social class (Currie et al. 1997; Wardle et al. 2002) and is conceptually related to common consumption indices of material deprivation (Carstairs and Morris 1990) and home affluence (Wardle et al. 2002). FAS items ask students about things they are likely to know about in their family (car, bedrooms, vacations, and computers), thus limiting the number of non-responses in the study. When the scale was introduced it was used in a national context only and contained three items (family car, bedroom and telephone) (Currie et al. 1997). In 2001/02 it was used cross-nationally and comprised family car, bedroom, holiday and computer. Within the HBSC study, the three-item scale (family car, bedroom and holiday) is named FAS I and the four-item scale (FAS I and computer) is named FAS II. The items, their response categories, and their rationale are the following:

- *Does your family own a car, van or truck?* (No = 0, Yes, one = 1, Yes, two or more = 2) This item is a component of the Scottish deprivation index developed by Carstairs and Morris (Carstairs and Morris 1990), which is used widely in health inequalities research.
- *Do you have your own bedroom for yourself?* (No = 0, Yes = 1) This item is a simple proxy for overcrowding, classified by Townsend (1987) as housing deprivation, and is also a component of the Scottish deprivation index.
- *During the past 12 months, how many times did you travel away on holiday with your family?* (Not at all = 0, Once = 1, Twice = 2, More than twice = 3) This item is a measure of 'deprivation of home facilities' (Townsend 1987).

- *How many computers does your family own?* (None = 0, One = 1, Two = 2, More than two = 3) This item has been introduced to differentiate SEP in affluent countries.

FAS has earlier been limited to including structural measures, based on family possessions. Data from the HBSC study contains an item assessing the student's perception of the family's wealth. The child is asked *How well off do you think your family is?* Responses are 1 = *Very well off*, 2 = *Quite well off*, 3 = *Average*, 4 = *Not very well off*, 5 = *Not at all well off*. The basis for including this item was the fundamental theoretical consideration that reliability increases when adding items. The analyses were done on the five item scale. The five items revised FAS II is referred to as FAS III in the following.

1.2 Comparative Studies on HBSC Data

In their study Torsheim and colleagues used FAS I scores aggregated to the level of school and country as indicators of contextual deprivation at these levels. The study found effects on self-rated health of deprivation on several levels, and that the most disadvantaged 11-year olds, were eight times more likely to have poor self-rated health compared to the least disadvantaged students (Torsheim et al. 2004). Elgar and colleagues transformed FAS I to cumulative probabilities and constructed a continuous material deprivation score ranging between 0 and 1, for use in multilevel logistic regression analyses on alcohol consumption. The study suggests that the distribution of wealth within societies may indirectly influence the use of alcohol during early adolescence (Elgar et al. 2005). Due and colleagues included FAS I as a confounder at the individual level in analysing relations between bullying and symptoms by multilevel logistic regression. This study found consistent and strong associations between bullying and symptoms in the 28 HBSC countries included (Due et al. 2005). Vereecken and colleagues used school and country averages of FAS II as second and third level variables. Within each country the individual level FAS was recoded into tertiles in analysing relations between social indicators and intake of fruit and soft drinks by multilevel logistic regression. This study suggested that socio-economic factors play a role in relation to the food habits of young adolescents, since it found that fruit intake increased with family material wealth and higher parental occupational status. The study also found that the intake of soft drinks was lower among more wealthy students in Northern, Southern and Western European countries, but higher among the wealthy students in Central and Eastern European countries (Vereecken et al. 2005). In a recently published paper from Torsheim and colleagues on the health effects of relative social inequality, the country distribution of family affluence was measured by the standard deviation of FAS I and compared across countries. This study used FAS I as a supplement to conventional income based indicators by using the standard deviations of FAS as a proxy for the within-country distribution of family material indicators. The study found substantial inequalities in subjective health across European and North American countries related to the distribution of family material resources in these countries (Torsheim et al. 2006).

The above mentioned indicate that FAS has been included differently depending on the research question, the data being analysed and the statistical technique in use. It also reflects that FAS has great potential to be included as an individual, school or country level variable.

1.3 Validity of FAS

Based on the thorough development of the survey items and the theoretical rationale of the scale, it is assumed that the scale is content valid (Currie et al. 1998) (pp.64–72). The scale has been used widely in national studies as an absolute measure, and some studies have included FAS in comparative absolute analyses. The possible DIF of FAS across countries has been previously examined by Batista-Foguet and colleagues, but using a different statistical approach (Batista-Foguet et al. 2006).

A critical question when relying on a scale derived from items presumed to be sensitive to an attribute (in this case family affluence) is the scale's measurement properties (Bjorner et al. 1998a). Collecting data according to a standardized protocol as done in HBSC allow researchers to compare data from different countries. However, a critical evaluation of the comparability of questionnaires across cultures and languages is only partly achieved by back translation, review by focus groups and co-ordinated multinational questionnaire development (Bjorner et al. 1998b). Statistical methods for translation evaluation are available, one of which is the analysis of DIF or item bias. DIF methods require that items should function in the same way, whatever subgroups are investigated. Examples of DIF in relation to gender have been observed in scales measuring physical ability among the elderly, Activities in Daily Living (ADL). Here DIF was observed in relation to gender in the question *Can you comb your hair?* This finding was interpreted that men and women relate to their hair differently, or simply that men at this age often does not have enough hair to have to comb it. DIF was also observed in relation to country in the question *Can you take a walk in bad weather?* Something other than the physical ability of the elderly was observed in item responses, which was that *bad weather* differs in different countries, which was reflected in item responses¹. Methods for investigating DIF can also be used to correct for DIF, depending on the fit of the model (Bjorner et al. 1998b).

Batista-Foguet and colleagues have found that there are country specific variations in the relative contribution of the three items of the FAS I, and their work provided a new strategy for weighting scale items by suggesting a Revised FAS. Their work concluded that cross-cultural studies were possible if constructing a country specific index (Batista-Foguet et al. 2006). Their study proposed an approach for performing comparative analysis by means of Adapted Canonical Variate Analysis (Batista-Foguet et al. 2006). This method, however, has not been widely used so far.

The aim of this paper is to study the FAS and from the perspective of comparative studies to examine whether it is comparable across population subgroups.

2 Method

The construct validity and objectivity of the FAS III was investigated by Rasch's model for item analysis and its generalization to Graphical Log Linear Rasch Models (GLLRM).

There are a number of technical demands, which need to be met when validating the measurement properties of a scale. The Rasch model is an item response model, which can be regarded as a formalization of perfect measurement, by meeting the requirements of criterion related *construct validity* (Rosenbaum 1989), statistical *sufficiency* (Andersen 1973) and *objectivity* (Rasch 1966). Rosenbaum's definition of construct validity is but one of many attempts to define construct validity. One particular important approach may be

¹ These examples are taken from discussions with colleagues at the Department of Social Medicine.

found in the seminal paper by Cronbach and Meehl (1955/2006). We refer to Borsboom (2005), Kane (2006) and Zumbo (2007) for discussions of validity. Kreiner (2007) discusses a family of Item Response Theory (IRT) models referred to as graphical IRT models in which the definitions of Rosenbaum and Cronbach & Meehl are reconciled.

Two of the requirements of construct validity according to Rosenbaum are that items are mutually locally independent (no local dependence, LD) and conditionally independent of exogenic variables given the latent variable being measured (no differential item functioning, DIF). Like Zumbo (2007; 56) we do not regard the question of validity as an all-or-none decision, but rather as a question of the degree of validity, although degree here is taken in a qualitative rather than a quantitative sense. The GLLRM relaxes these two requirements, replacing them with the requirements that DIF and LD are uniform in the sense that strength of association between locally dependent items, and the effect of exogenic variables on items does not depend on the trait being measured. Measurement by items from GLLRM are therefore not construct valid according to Rosenbaum's definition, but Kreiner and Christensen (2006) and Kreiner (2007) claim that items from a GLLRM provide measurement, which are *essentially* valid and objective. Uniform local dependence in itself is not a problem for the validity, although reliability may be affected, and test equating techniques may be used to adjust for the confounding effect of uniform DIF.

Inference in GLLRM is discussed in Kreiner and Christensen (2002, 2004, 2006). Conditional inference based on the conditional distribution of item responses may be performed the same way for GLLRM as for conventional Rasch models. This is done given the total person scores in order to avoid assumptions concerning the distribution of latent traits. DIF and LD may be tested using both Mantel–Haenszel techniques and log linear tests as described by Kelderman (1984). Andersen's conditional likelihood ratio tests of fit (Andersen 1973) may also be calculated for GLLRM as many conventional item fit statistics like Molenaar's U and in-and outfit statistics comparing between observed and expected item response curves. Marginal inference based on assumptions relating to the distribution of the latent variable is also feasible, but will be more complicated than in ordinary item response models, because the assumptions have to relate to the conditional distribution of the latent variable given the exogenous covariates in the model.

Being able to include FAS in analyses across countries is of particular interest for the purpose of this study, and comparisons across countries may be carried out in the presence of DIF if DIF and LD are uniform. However, if item bias is not uniform, cross-country comparisons may be spurious.

2.1 DIF Equating

Assume that person n in group g has a total score equal to S_{gn} and that person m from group h has a total score equal to S_{hm} and that some of the items but not all function differentially in the two groups. The two scores are obviously not comparable, because of the differential item functioning, but the two estimates of the values of the latent variable, θ_{gn} and θ_{hm} are comparable because both estimates are values on the same latent scale. Now, let $T_{gn} = E(S|\theta_{gn}, \text{Group} = 0)$ $T_{hm} = E(S|\theta_{hm}, \text{Group} = 0)$. We refer to T_{gn} and T_{hm} as equated scores, because the transformation of values on the latent scale to expected values of the manifest score is one of the standard techniques used for test equating. Instead of comparing the values on the latent scale one may compare the equated scores. Comparison of equated scores is, of course, neither better nor worse than comparison of the values of the latent FAS variable, but it may be preferable if one feels

more comfortable with the manifest FAS scores than with the corresponding estimates of the latent FAS values.

3 Results

3.1 Progress of Analyses

Three types of analyses were performed. First, an item analysis attempting to identify and fit one common GLLRM to all countries was performed. For such an attempt to succeed there has to be at least one polytomous (or unbiased “anchor”) item and preferably more items that do not function differentially relative to country. This attempt failed for which reason further attempts to analyse the complete data set was abandoned.

Secondly, GLLRM were fitted to each country separately. The results of these analyses are presented below, together with an attempt to describe the common features of the different country specific models.

The success of the country specific analyses means that FAS measurements may be used for comparisons and analyses within countries, but does not support between country comparisons and analyses for all countries. The failure of the attempt to fit an item response model does not necessarily mean that such applications of FAS is impossible, e.g. consider a case with 5 items and 3 countries. The failure to identify one item with no DIF relative to country may be because items 1 to 2 function differently in Country 1 and Country 2, while the remaining items function differently in Country 2 and 3. In this case, analysis including only Country 1 and 2 would identify one GLLRM that might be used to equate scores from Country 1 to Country 2, while another GLLRM could be used to equate scores from Country 3 to Country 2. If both these analyses succeeded combining results from the two analyses could be used to equate scores from Country 1 to Country 3.

The final analyses attempted to link countries by analyses of data from pairs of countries. UK was used as the reference country in all of these analyses, since the UK sample was larger than the samples from the other countries, since UK was a combination of England, Scotland and Wales. The analyses failed to identify a GLLRM for all countries, but succeeded in enough cases for between country comparisons to be of interest.

3.2 Reliability

To test the reliability of the different FAS scales, we calculated Cronbach’s α for the FAS II and FAS III. The results of these analyses showed that Cronbach’s α for FAS II varied between countries from 0.20 to 0.60 and for FAS III, it varied between 0.32 and 0.62 (data not shown). The highest increase observed was 0.12, and the average increase was 0.06. The item *perceived family wealth* showed great consistency and adapted to the latent variable expressed by FAS II on the same terms as the existing four items. The inclusion of the fifth item thus increased reliability.

3.3 Conditional Likelihood Ratio test

Initial analyses examined the DIF in relation to age, gender and country, and also in relation to different levels of FAS III score. Table 1 shows the results of conditional

likelihood ratio (CLR) tests, comparing item parameter estimates in the four different groups. The highly significant P -values reveal that DIF is present in relation to age, gender and country. Comparison of estimates of item parameters in different score groups showed that item responses were heterogeneous relative to the latent level of FAS III (Table 1). All CLR tests are highly significant, but the CLR test relating to country is particularly noteworthy. Attempt to fit GLLRM failed because of the complex DIF effects relative to country, meaning that models could not be identified. The country related DIF therefore appeared to obstruct cross country comparative analyses on the basis of FAS III as it stands.

3.4 The Fitting of Graphical Log-Linear Rasch Models

Fitting GLLRM's and test equating for DIF effects require that there is at least one item with no DIF relative to each of the exogenous variables. The failure of GLLRM for the data set with all countries may be the result of a very complex DIF structure relative to country where no item (or variable) functions the same way in all countries. GLLRM may, however, fit pairs of countries such that test equating between these countries is possible. For this reason, the last analysis fitted GLLRM to all countries with data from the UK, and constructed adjusted scores by test equating. This was done with the youngest boys in the UK as the reference group. The consequence of this is mentioned later in this section (in *adjusted scores*).

Due to the above mentioned reasons, the fit of FAS III to GLLRM was initially performed for each of the 32 countries. Estimates of GLLRM parameters and *partial gamma coefficients* (denoted γ^p) measuring the strength of the conditional association between pairs of items, and between items and exogenous variables were used for analysis of DIF and LD. In this paper, γ^p correlations of 0.00–0.15 were regarded as weak correlations, 0.16–0.30 were regarded as moderate correlations and $\gamma^p > 0.30$ were regarded as strong correlations. Table 2 provides with an overview of the γ^p calculated on the basis of 32 analyses (Table 2).

Generally, the table shows patterns in whether the correlation is positive or negative, with only a few exceptions, which can be considered type 1 errors.

Due to space requirements for this paper, two models were selected to be presented as examples of the 32 sets of analyses performed, these where the models of Denmark and the United Kingdom. The GLLRM model fitting item responses for Denmark is shown in Fig. 1. Figure 2 illustrate the findings for the UK, which was selected a pairwise reference region, due to the large sample size.

Evidence of DIF was found between age and three items: bedroom, holiday and well off. Conditionally given the true FAS level, the findings in Denmark where that the chances of having your own bedroom increased with age ($\gamma^p = 0.32$) while the chances of

Table 1 Conditional likelihood ratio test comparing item parameter estimates in different groups

	CLR	df	P -value
Score groups	3,226.2	8	0.000
Age	4,184.3	16	0.000
Gender	479	8	0.000
Country	85,552.2	248	0.000

Data from the international HBSC survey 2001/02

Table 2 continued

Country	Differential item functioning (DIF)										Local dependency (LD)									
	Room/ Gnd	Room/ Age	Car/ Gnd	Car/ Age	Hol/ Gnd	Hol/ Age	W.O./ Gnd	W.O./ Age	Comp/ Gnd	Comp/ Age	Room/ Car	Room/ Hol	Room/ W.O.	Room/ Comp.	Car/ W.O.	Car/ Hol	Car/ Comp.	Hol/ W.O.	Hol/ Comp.	W.O./ Comp.
Sweden	-	0.44	-	0.14	0.17	-	-	-0.38	-0.21	0.10	0.22	-	-	-	-	-	0.14	-	-	-
Switzerland	-	0.29	0.16	-	-	-0.17	-	-0.13	-	0.12	-	-	-	0.21	-	-	0.10	-	-	-
Ukraine	-	0.16	0.16	-	0.07	-	-	-0.41	-0.26	0.08	-	-	-	-	-	0.03	0.21	-0.10	-0.04	0.16
Unit. Kingd.	-	0.20	-	-	-	-0.15	-	-0.03	-0.18	0.01	0.10	-	-	-0.09	-0.04	0.02	0.11	0.07	-0.06	-0.07
USA	-	0.25	-	-	-	-0.11	-	-0.15	-	-	0.11	-	-	0.05	-0.11	0.04	0.25	-	-	-0.18
No. countries where evidence of DIF/LD is disclosed	8	25	5	12	7	22	8	27	17	16	11	2	2	8	5	11	30	18	5	8

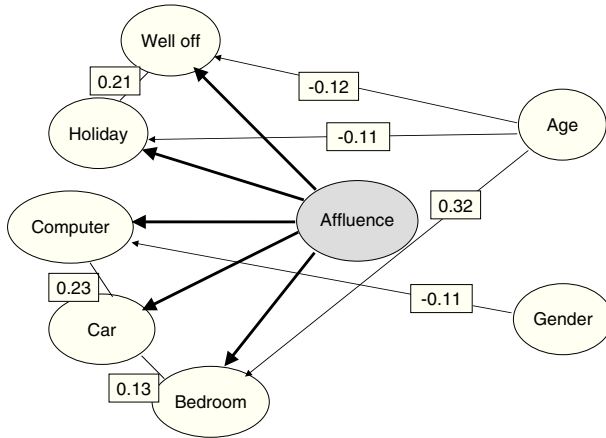


Fig. 1 The generalization of FAS III to GLLRM illustrated with the example of Denmark (number in boxes are partial gamma coefficients, γ^p , indicate a significant correlation beyond the latent trait)

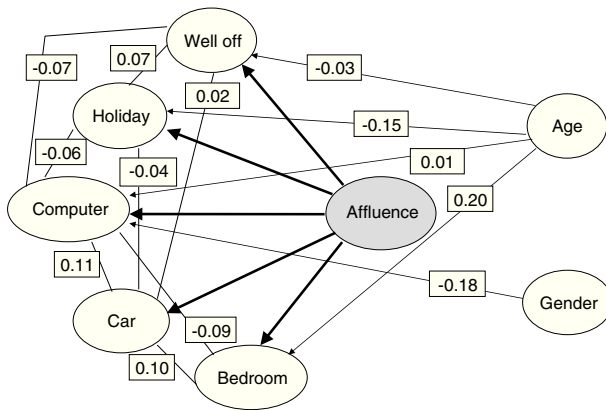


Fig. 2 The generalization of FAS III to GLLRM illustrated with the example of the UK (number in boxes are partial gamma coefficients, γ^p , indicates a significant correlation beyond the latent trait)

going on holiday with your family ($\gamma^p = -0.11$), and feeling that the family was well off decreased with age ($\gamma^p = -0.12$). Furthermore, in Denmark DIF was observed between gender and computers, indicating that families of boys—everything else being equal—more often had computers ($\gamma^p = -0.11$).

In the model fit for Denmark, three coefficients were significant revealing LD between items, and these were weak (bedroom \leftrightarrow family car) or moderate (family car \leftrightarrow computer, holiday \leftrightarrow well off). This meant that a positive evaluation of the level of family wealth was locally dependent of going on holiday ($\gamma^p = 0.21$), and there were correlation between the family having a car and the child having his/her own bedroom ($\gamma^p = 0.13$), and also between family car and computer ($\gamma^p = 0.23$).

In the model fit for the UK, an additional DIF was found (age \leftrightarrow computer) even if this was weak ($\gamma^p = 0.01$). One DIF was the same (age \leftrightarrow holiday), and two where weaker

(age \leftrightarrow bedroom, age \leftrightarrow well off). Evidence of LD with moderate γ^p was found for all relations from computer, which was only evident in one relation for the model fit for Denmark (family car \leftrightarrow computer). Two DIF relations showed weaker associations (age \leftrightarrow bedroom, age \leftrightarrow well off) for UK when comparing with Denmark, and one was the same (age \leftrightarrow holiday). One LD relation was weaker in the model for the UK (family car \leftrightarrow computer), and one (family car \leftrightarrow bedroom) was the same.

When summing up all the results from the 32 analyses, three correlations were considered general. What was observed in 12 countries was DIF between age and family car, meaning that the family more often had a car when the children were older. Another general finding was DIF between age and computers, showing that the families with older children more often had computers, independent of family wealth. The third correlation which was general, but not found in the model fit for Denmark, was the evidence of LD in the relation between family car and well off, since ownership of a car was related to considering the family well off.

Table 2 reveals the evidence of DIF disclosed throughout the 32 countries. The analyses revealed DIF with respect to age in relation to bedroom in 25 of the 32 countries, in relation to family car in 12 countries, in relation to holiday in 22 countries, in relation to computer in 15 countries and in relation to being well off, in 27 of the 32 countries. When summing up the evidence of LD found in each of the 32 countries, the analyses revealed LD between family car and bedroom in 11 of the 32 countries, between family car and well off in 11 countries, between holiday and well off in 18 countries, and between family car and computer in 30 of the 32 countries (bottom of Table 2).

Failure to fit a GLLRM to the complete set of data means that measurement may not be valid for comparisons between countries. However, it does not necessarily mean that mean that within country measurement is invalid.

3.5 Adjusted Scores

The equating techniques to adjust for the DIF were provided by the GLLRM analysis. Table 2 revealed general patterns in the DIF and LD across countries, and these led to the final analysis, which fitted GLLRM to all countries with data from the UK, and constructed adjusted scores by test equating techniques.

Table 3 summarizes the attempt to equate scores from different countries to UK scores. DIF equating between UK and another country requires at least one polytomous item that function in the same way in the two countries. Such a link was identified in 19 countries, but was not found for the remaining 13 countries included in this study.

Table 4 shows equated scores for UK and Denmark with 11 years old boys from UK as reference. As shown in Table 4, all values are slightly modified, and the consequences of the DIF with respects to age and gender become evident. A boy age 15 in Denmark has a score of 8.60 to correspond to a score of 7 for an 11-year old boy in the UK. Tables with DIF equated scores for the 19 countries where DIF equating to UK was possible is available from the HBSC Databank Manager at University of Bergen.

The adjusted scores are generally higher for girls than for boys at the same age, while adjusted scores are decreasing with age when observed scores are low, and increasing with age when observed scores are high. Note that the choice of reference group is arbitrary, and that adjustment of scores due to age and gender related DIF is limited. Adjustments in other countries may be more pronounced.

Table 3 Unbiased anchor items for DIF equating FAS III scores to UK data

Country	Unbiased anchor
Austria	Computer
Belgium	Holiday
Canada	Computer
Croatia	<i>No unbiased items</i>
Czech Republic	<i>No unbiased items</i>
Denmark	Computer
Estonia	Computer
Finland	Computer
France	Family car
Germany	Well off
Greece	<i>No unbiased items</i>
Greenland	Bedroom and holiday
Hungary	Family car and bedroom
Ireland	Family car
Israel	<i>No unbiased items</i>
Italy	Family car
Latvia	Family car ^a
Lithuania	<i>No unbiased items</i>
Macedonia	Bedroom and holiday
Malta	<i>No unbiased items</i>
Netherlands	<i>No unbiased items</i>
Norway	Computer
Poland	Computer
Portugal	Well off
Russia	Holiday
Slovenia	<i>No unbiased items</i>
Spain	<i>No unbiased items</i>
Sweden	Computer
Switzerland	Bedroom and computer
Ukraine	<i>No unbiased items</i>
USA	<i>No unbiased items</i>

^a Despite anchor item, this model showed very bad fit
Data from the international HBSC survey 2001/02

4 Discussion

This study confirmed prior findings by Batista-Foguet et al., that there are cross country variations in the relative contribution of the four items to the overall FAS. The large CLR value seen in Table 1 revealing DIF in relation to country, indicated that the model had a very poor fit, which led to analysis to detect DIF and LD performed at a country level.

That the scale performs differently in different countries is not surprising as prices of cars and housing can vary significantly depending on a country's political and economical structure. Furthermore, housing space can be considered a sociological issue concerning family size, and also how the family divides the housing between the members of the household.

Table 4 DIF equated scores, Denmark and the UK as examples

	Score	Boys			Girls		
		11	13	15	11	13	15
Denmark	1	1.17	1.13	1.14	1.16	1.13	1.14
	2	1.88	1.71	1.74	1.88	1.71	1.74
	3	3.21	2.89	3.00	3.27	2.94	3.05
	4	4.70	4.37	4.59	4.82	4.49	4.71
	5	6.08	5.88	6.16	6.24	6.03	6.32
	6	7.31	7.24	7.54	7.46	7.40	7.70
	7	8.34	8.37	8.60	8.49	8.51	8.73
	8	9.13	9.18	9.31	9.23	9.28	9.39
	9	9.67	9.70	9.74	9.72	9.75	9.78
	Score	Boys		Girls			
		11 (ref)	13	15	11	13	15
United Kingdom	1	1.00	1.21	1.22	1.25	1.21	1.22
	2	2.00	1.86	1.92	2.02	1.87	1.94
	3	3.00	2.78	2.93	3.08	2.84	3.00
	4	4.00	3.77	4.00	4.14	3.89	4.12
	5	5.00	4.82	5.10	5.18	4.99	5.27
	6	6.00	5.91	6.21	6.20	6.11	6.41
	7	7.00	7.00	7.28	7.21	7.22	7.50
	8	8.00	8.05	8.29	8.20	8.25	8.47
	9	9.00	9.05	9.19	9.14	9.18	9.31

This study found age-related DIF for the majority of the HBSC countries in relation to all items included in the scale. This problem would not be solved by adjusting for age as a confounder, since this modification does not adjust on an item level, but on the sum score. Older children more often had their own bedroom, independent of family wealth. These adolescents tended to rate their family as less affluent, and were also less likely to go on holiday with their parents. Furthermore, the study showed that the older children more often had computers, and also the somewhat unexpected but clear finding, that families with boys more often had computers in their home. LD was identified between well off and the items car and holiday. These three items seemed to be correlated beyond family affluence. The positive evaluation of the family's wealth was dependent on holiday and car ownership. Car and computer ownership was correlated in all countries except for two, which indicated that these two items do not differentiate the scale much.

On the basis of the GLLRM, converted scores are constructed for use in 19 of the 32 countries. For the remaining 13 countries, comparative studies using FAS III is not advisable, due to DIF with respect to country, age, and gender.

If one item had not shown DIF in relation to country, this could have been regarded as an "anchor item", and the scale could be recoded with this item as a "fix point". However, this was not the case in this study, and the consequence is that the use of absolute sum scores in international comparisons may result in spurious outcomes.

A cautious approach to the use of the DIF equated scores from different countries is wise. For most of the countries it was only possible to identify one item that functioned in the same way in as in the UK. This is sufficient for calculation of DIF equated scores, but it is usually not regarded as enough evidence supporting claims that one is actually measuring the same construct in the different groups. Development of additional FAS items, some of which functions in the same way between countries, is consequently of utmost importance in order to support the claims that DIF equating of FAS scores result in valid comparisons between countries.

This study recommends using the adjusted FASIII measure to adjust for the presence of DIF but cannot indicate how large an impact this will have on results. Empirical analysis is required to investigate the effect of adjusting for DIF on the outcome of analyses. One limitation of the findings in this study is that the DIF adjusted FAS may only be relevant in studies using HBSC data. In studies where adjustment of FAS is not pragmatic e.g. for countries not included in the HBSC, it is useful to know the impact of using adjusted FAS versus unadjusted FAS on outcomes of interest, and whether such differences can be generalized across countries. Furthermore, the impact of adjusting for the magnitude of DIF should be assessed independently of the impact of adjusting for the direction of DIF.

The study concludes that FAS can be used as an interval scale and an absolute as well as a relative measure of wealth within a country and between countries, provided that the item on perceived family wealth is included, and the scale is converted to adjust for DIF. The effects on health and other outcomes of this relative measure of social position can be compared across countries, and if following the above mentioned steps the validity of FAS will increase. Using this method the HBSC survey is a potential valuable data source for comparative studies on health inequalities among school children.

4.1 Methodological Considerations

In our experience, DIF and LD appear to be the rule rather than the exception in health related scales. GLLRM are useful in such circumstances and they are relatively simple to work with, because conditional inference and Mantel–Haenszel techniques apply for these models in exactly the same way as for ordinary Rasch models. The models are, however, restricted to situations where sufficiency applies and DIF and LD are uniform. More general IRT models where item discriminations differ across items and where DIF and LD are not uniform are feasible, but software for such models are not available at the moment. In these cases one has to fall back on the old strategy of sacrificing items even though they might provide essentially valid and more reliable measurements.

We validate a scale to make sure that measurements are not confounded by DIF and/or items measuring something else, but in most cases, the problem of validating a scale is subordinate to the purpose for which the scale was included in the study. Analyses addressing the dependence of the *latent* variable, or the way the latent variable influences or is associated to other variables, is referred to as *latent regression analyses* or *latent structure analyses*. GLLRM are useful for such purposes, and preferable to analyses where the score—DIF equated or not—is used as a proxy for the latent variable. Latent regression in GLLRM has been described by Christensen and colleagues (2004). SAS macros for GLLRM have been developed by Christensen and Bjørner (2003). Compared to these methods, analyses using scores as proxies for latent variables is only the second best solution, because relationships that are linear in terms of the latent variable in most cases will not be linear in terms of the score.

Eventually, we wish to remind the reader that the kind of DIF equating described in this paper requires unidimensionality and uniform DIF. It is inappropriate if DIF is non-uniform. In situations where DIF is found, we usually argue that it must be the smaller subset of items that function differentially, which nevertheless does not follow automatically. A cautious view on the DIF phenomenon is therefore prudent, in particular in scales with few items.

Acknowledgements The HBSC study is an international study carried out in collaboration with WHO Europe. The international coordinator is Professor Candace Currie from the University of Edinburgh, and the international databank manager is Dr. Oddrun Samdal from Bergen University. The 2001/02 survey was conducted by personal responsibility of principal investigators (PI's) in 35 countries. The authors are grateful for the insightful and constructive comments to earlier drafts from Ms. Kate Levin and Ms. Dorothy Currie from the University of Edinburgh, and also for constructive feedback late in the process from the PI's.

References

- Andersen, E. B. (1973) A goodness of fit test for the Rasch model. *Psychometrika*, *38*, 123–140.
- Batista-Foguet, J. M., Fortiana, J., Currie, C., & Villalbi, J. R. (2006) Socio-economic indexes in surveys for comparisons between countries. *Social Indicators Research*, *67*, 315–332.
- Bjorner, J. B., Damsgaard, M. T., Watt, T., & Groenvold, M. (1998a) Tests of data quality, scaling assumptions, and reliability of the Danish SF-36. *Journal of Clinical Epidemiology*, *51*, 1001–1011.
- Bjorner, J. B., Kreiner, S., Ware, J. E., Damsgaard, M. T., & Bech, P. (1998b) Differential item functioning in the Danish translation of the SF-36. *Journal of Clinical Epidemiology*, *51*, 1189–1202.
- Borsboom, D. (2005). *Measuring the mind*. United Kingdom: Cambridge University Press.
- Carstairs, V., & Morris, R. (1990) Deprivation and health in Scotland. *Health Bulletin (Edinburgh)*, *48*, 162–175.
- Christensen, K. B., Bjorner, J. (2003) Sas macros for Rasch based latent variable modelling. Technical report 03/13. Department of Biostatistics, University of Copenhagen.
- Christensen, K. B., Bjorner, J., Kreiner, S., & Petersen, J. H. (2004). Latent regression in log linear Rasch models. *Communications in Statistics*, *33*, 1295–1314.
- Cronbach, L. J., & Meehl, P. E. (1955/2006). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302 (Reprinted in Bartholomew, D. J. (Ed.) (2006). *Measurement. Vol. 1* (pp. 277–306). London, Sage Publications).
- Currie, C. et al. (1998). Health behaviour in school-aged children: a WHO cross-national survey (HBSC), Research Protocol for the 1997/98 Survey. Research Unit in Health and Behavioural Change. 1998. Edinburgh, World Health Organisation. Ref Type: Report.
- Currie, C. E., Elton, R. A., Todd, J., & Platt, S. (1997) Indicators of socioeconomic status for adolescents: the WHO Health Behaviour in School-aged Children Survey. *Health Education Research*, *12*, 385–397.
- Due, P., Holstein, B. E., Lynch, J., Diderichsen, F., Gabhain, S. N., Scheidt, P., & Currie, C. (2005) Bullying and symptoms among school-aged children: international comparative cross sectional study in 28 countries. *European Journal of Public Health*, *15*, 128–132.
- Elgar, F. J., Roberts, C., Parry-Langdon, N., & Boyce, W. (2005) Income inequality and alcohol use: a multilevel analysis of drinking and drunkenness in adolescents in 34 countries. *European Journal of Public Health*, *15*, 245–250.
- Kane, M. T. (2006) Validation (p.17–65). In R. L. Brennan (Eds.), *Educational measurement*, 4th edn. Westport: Praeger Publishers.
- Kelderman, H. (1984) Loglinear Rasch Tests 49. *Psychometrika*, *49*, 223–245.
- Kreiner, S., & Christensen, K. B. (2002) Graphical Rasch models. In M. Mesbah, B. F. Cole, & M. T. Lee (Eds.), *Statistical methods for quality of life studies. Design, measurements and analysis*. Dordrecht: Kluwer Academic Publishers, pp. 187–203.
- Kreiner, S., & Christensen, K. B. (2004) Analysis of local dependence and multidimensionality in graphical loglinear Rasch models. *Communications in Statistics— Theory and Methods*, *33*, 1239–1276.
- Kreiner, S., & Christensen, K. B. (2006) Validity and objectivity in health related summated scales: Analysis by graphical log-linear Rasch models. In M. von Davier, C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models. Extensions and applications*. New York: Springer Verlag.

- Kreiner, S. (2007) Validity and objectivity. Reflections on the role and nature of Rasch models. *Nordic Psychology*, *59*, 268–298.
- Lien, N., Friestad, C., & Klepp, K. I. (2001) Adolescents' proxy reports of parents' socioeconomic status: How valid are they? *Journal of Epidemiology and Community Health*, *55*, 731–737.
- Mackenbach, J. P., & Bakker, M. J. (2002) *Reducing inequalities in health: A European perspective*. London: Routledge.
- Rasch, G. (1966) An item analysis which takes individual differences into account. *The British Journal of Mathematical and Statistical Psychology*, *19*, 49–57.
- Rosenbaum, P. R. (1989) Criterion-related construct validity. *Psychometrika*, *54*, 625–633.
- Torsheim, T., Currie, C., Boyce, W., Kalnins, I., Overpeck, M., & Haugland, S. (2004) Material deprivation and self-rated health: a multilevel study of adolescents from 22 European and North American countries. *Social Science and Medicine*, *59*, 1–12.
- Torsheim, T., Currie, C., Boyce, W., & Samdal, O. (2006) Country material distribution and adolescents' perceived health: multilevel study of adolescents in 27 countries. *Journal of Epidemiology and Community Health*, *60*, 156–161.
- Townsend, P. (1987) Deprivation. *Journal of Social Policy*, *16*, 125–146.
- Vereecken, C. A., Inchley, J., Subramanian, S. V., Hublet, A., & Maes, L. (2005) The relative influence of individual and contextual socio-economic status on consumption of fruit and soft drinks among adolescents in Europe. *European Journal of Public Health*, *15*, 224–232.
- Wardle, J., Robb, K., & Johnson, F. (2002) Assessing socioeconomic status in adolescents: the validity of a home affluence scale. *Journal of Epidemiology and Community Health*, *56*, 595–599.
- Whitehead, M., Diderichsen, F., & Burstrom, B. (2000) Researching the impact of public policy on inequalities in health. In H. Graham (Ed.), *Understanding health inequalities*. London: Open University Press.
- Zumbo, B. D. (2007) Validity: Foundational issues and statistical methodology. In: C. R. Rao, & S. Sinharay (Eds.), *Handbook of Statistics, volume 26: Psychometrics*. Amsterdam: Elsevier.