

Equating the HBSC Family Affluence Scale across survey years: a method to account for item parameter drift using the Rasch model

Guido Makransky · Christina Warrer Schnohr ·
Torbjørn Torsheim · Candace Currie

Accepted: 2 December 2013 / Published online: 6 June 2014
© Springer International Publishing Switzerland 2014

Abstract

Purpose To investigate the measurement invariance (MI) of the Family Affluence Scale (FAS) measured in the Health Behavior in School-aged Children (HBSC) survey, and to describe a method for equating the scale when MI is violated across survey years.

Methods This study used a sample of 14,076 Norwegian and 17,365 Scottish adolescents from the 2002, 2006 and 2010 HBSC surveys to investigate the MI of the FAS across survey years. Violations of MI in the form of differential item functioning (DIF) due to item parameter drift (IPD) were modeled within the Rasch framework to ensure that the FAS scores from different survey years remain comparable.

Results The results indicate that the FAS is upwardly biased due to IPD in the computer item across survey years in the Norwegian and Scottish samples. Ignoring IPD across survey years resulted in the conclusion that family affluence is increasing quite consistently in Norway and Scotland. However, the results show that a large part of the increase in the FAS scores can be attributed to bias in the FAS because of IPD across time. The increase in the FAS was more modest in Scotland and slightly negative in Norway once the DIF in the computer item was accounted for in this study.

Conclusions When the comparison of family affluence is necessary over different HBSC survey years or when the longitudinal implications of family affluence are of interest, it is necessary to account for IPD in interpretation of changes in family affluence across time.

Keywords Family Affluence Scale · Differential item functioning (DIF) · Health Behavior in School-aged Children (HBSC) · Rasch model · Item parameter drift (IPD)

Introduction

Socioeconomic status (SES) is consistently related to a number of health-related outcomes; therefore, variations in social inequalities in health are a fundamental part of social epidemiological research [1]. Although most of this research has been done with adult populations, reducing inequalities in child and adolescent health are essential societal interventions that can improve the conditions and the general health of future generations. A valid and comparable measure of SES is essential for understanding health-related outcomes nationally and internationally; however, questionnaires to adolescents on their parents' education, occupation or income usually result in low completion rates and a high misclassification rate [2–5], so alternatives to the common proxies of SES have been investigated.

The Health Behavior in School-aged Children (HBSC) is an international study with 43 participating countries and regions in the European Region and North America. The study aims at providing comparable data on young people's health and lifestyle from countries with different economic conditions and cultural, societal and political systems. In

G. Makransky (✉)
University of Southern Denmark, Odense, Denmark
e-mail: gmakransky@health.sdu.dk

C. W. Schnohr
University of Copenhagen, Copenhagen, Denmark

T. Torsheim
University of Bergen, Bergen, Norway

C. Currie
University of St. Andrews, St. Andrews, Scotland

the HBSC, SES has been investigated as family affluence measured by collecting information on material assets, in the Family Affluence Scale (FAS). First developed in a national context [3], the FAS was subsequently included in the HBSC international study from 1998 comprising three items and four items since 2002 [6]. The scale is widely used for presenting trend analyses across survey years and guiding national policy decisions. From its early development, the validity of the FAS has been discussed in several papers and validated at both national and international levels. Studies found that the FAS has good criterion validity, on the basis of showing graded associations between socioeconomic position (SEP) and various health outcomes [7–9]. Adolescents and parents report similarly to the FAS items [10], and the FAS is less affected by non-response bias than SEP measures that rely on child reports of household income or parental occupation [3, 6, 11]. Additionally, data from the FAS are far less burdensome to collect and manage than other sources on SEP, e.g., data on parental occupation. Within educational research, the FAS is considered valuable as a tool when assessing children's eligibility for free school meals, but in general, it has been concluded that the FAS does not have good overall reliability [5]. The low reliability is mainly a consequence of the circumstance that the scale consists of as few as four items.

The traditional approach to using the FAS is to treat the items as indicators of the underlying family affluence (latent) construct and to add up the items to produce a sum score [2, 12]. The property of item invariance must be met in order for the sum score to be valid [13]. Item invariance requires that item estimation be independent of the subgroups of individuals completing the measure. In other words, item parameters have to be invariant across populations or time points [14]. The use of the sum score can be problematic when the FAS is used across survey years because the statistical properties of an item can change over time. Items that do not demonstrate invariance are commonly referred to as exhibiting differential item functioning (DIF). DIF occurs in the current context when individuals at different time points (e.g., 2002, 2006 and 2010 HBSC surveys) have different scores on specific items despite having equal levels of family affluence. This phenomenon is denoted as item parameter drift (IPD) [15, 16].

IPD is specifically relevant for the FAS because the items are direct measures of material assets. The prevalence of the number of computers in a household, the number of holidays per year or the ownership of a car which are items in the FAS can vary across time. Trends such as the increased access to personal computers, the general tendency that airline tickets have become less expensive or laws and market conditions that can make

cars more or less affordable can introduce IPD because these factors could increase the mean scores of the individual items, but may not be an indication of a general increase in family affluence over time. A valid and comparable measure of family affluence across survey years requires evidence of measurement invariance (MI) or the absence of DIF across time [17]. When MI can not be established across survey years, methods are needed that can account for IPD in order to obtain measures that are comparable across time [18].

Although there have been a number of studies that have identified the problematic nature of using the FAS to make comparisons across countries and time points [19, 20], the field is still missing a simple equating mechanism that can be used to make the measurement of the FAS across different HBSC survey years comparable.

Equating the FAS scores across survey years

The property of MI is met when items fit a Rasch model [21]. The Rasch model [13] can thus be used to test for MI by investigating if there are items that exhibit DIF across time points. The differences across time points for items with DIF can then be modeled by assigning group-specific item parameters [18, 22–24]. The method is a common items non-equivalent group's design where there are a number of common anchor items that do not exhibit DIF across time points. These items are used as an anchor test that provides the basis for comparison across time points in the Rasch model. The items that exhibit DIF across time points are then treated as virtual items, one for each time point. The items that exhibit DIF can then be modeled so that the location is empirically estimated at each time point within a calibration sample. In this way, the anchor items have common parameters across all time points, and the DIF items have time point-specific item parameters. This approach can only be used if it can be clearly shown that the responses to the items at the different time points apply to the same latent construct. In other words, the construct that is being measured must remain the same at different time points. This can be shown by investigating if the items fit the Rasch model for the entire set of response data [18, 25]. The reasoning behind this model is that items can have slightly different true parameters across time points. These differences can be modeled when there is statistical evidence to support the hypothesis that the items measure the same construct across the conditions.

In the current study, we illustrate a straightforward method within the Rasch model framework that can be used to assess DIF across survey years and account for IPD to ensure that the FAS scores from different survey years remain comparable. Furthermore, we provide examples of how the method works and assess the implications of using

this method with empirical data from the HBSC survey in 2002, 2006 and 2010 in Norway and Scotland.

Method

Measures

The FAS items ask students about things they are likely to know about in their family, thus limiting the number of non-responses in the study. When the scale was introduced in 1998, it was used in a national context only and contained three items (family car, bedroom and telephone) [3]. In 2001/2002, it was used cross-nationally and comprised family car, bedroom, holiday and computer. The items, their response categories and their rationale are the following:

1. *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 [2], which is used widely in health inequalities research.
2. *Do you have your own bedroom for yourself?* (No = 0, Yes = 1). This item is a simple proxy for overcrowding, classified by Townsend [12] as housing deprivation, and is also a component of the Scottish deprivation index.
3. *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” [12].
4. *How many computers does your family own?* (None = 0, One = 1, Two = 2, More than two = 3). This item was introduced to differentiate SEP in affluent countries.

Participants

Data from the FAS administered in the HBSC in years 2001/2002, 2005/2006 and 2009/2010 for Norway and Scotland were used. The Norwegian sample consisted of 14,076 adolescents, 7,167 boys and 6,909 girls, in age groups of 11, 13, 15 and 16 (Mean = 13.44). The Scottish sample consisted of 17,365 adolescents, 8,648 boys and 8,717 girls, between the ages of 10 and 16 (Mean = 13.57).

Statistical analyses

The Rasch model provides a solution to modeling DIF that occurs as a consequence of IPD across time points. When a scale is fitted to a Rasch model, raw scores are obtained for an item or person providing statistical sufficiency [13]. In other words, when able to fit a Rasch model, the person total score contains all information available within the specified context about the individual, and the item total

score contains all information with respect to the item, with regard to the relevant latent trait. Conversely, if a Rasch model cannot be fitted, it may be problematic to use a sum score because this score includes measurement bias which in turn can bias the interpretation of the results. The fundamental concept of the Rasch model is that each test item is characterized by a number of location or threshold parameters based on the number of response categories in the item, and each respondent is characterized by a single trait-level parameter. The probability that a given respondent answers within a certain response category for an item is given by a function of both the items and the parameters of the respondents. Conditional on those parameters, the response on one item is independent of the responses to other items. The Rasch model has the following form [13],

$$P_i(\theta) = \frac{\exp^{(\theta - b_i)}}{1 + \exp^{(\theta - b_i)}} \quad (1)$$

Here $P_i(\theta)$ is the probability of a certain response for item i , θ is the respondent's trait level and b_i represents the item location parameters. An extension of the Rasch model to items with more than two response options (polytomous items), the partial credit model (PCM) [26], is applied in this study.

The first step in modeling DIF is to use model-fit statistics to identify the items with DIF. The lagrange multiplier (LM) statistic is one method used in the assessment and modeling of DIF [18, 23, 24]. The LM statistic is a general tool for the evaluation of fit to IRT models, and in addition to the evaluation of DIF, it can also be used for evaluation of other assumptions of IRT, such as the form of the response curves and local independence [25, 27]. The LM test was chosen in this study because it can be used to test general model fit once DIF has been modeled with group-specific item parameters, in addition to identifying items with DIF across groups [18]. The general process of identifying and modeling DIF with the LM test is as follows:

The sample of respondents is divided into subsamples from subpopulations. These are the three survey year samples used to investigate DIF and the score-level groups used to evaluate model fit within subpopulations. Then, the item parameters are estimated. The item with the largest value on the LM test statistic targeted at DIF is then identified. The statistic is based on the difference between average observed scores on every item in the subsamples. The hypothesis tested is equivalent to testing the hypothesis that the parameters of the items are equal for the subgroups (for more details see [25, 27, 28]).

The LM statistic has an asymptotic chi-squared distribution that results in a significance test. However, such a test is only relevant for moderate sample sizes; for large sample sizes, the test becomes less interesting because its power becomes so large that even the

smallest deviations from the model become significant [28, 29]. The statistic can therefore be accompanied by the effect sizes which show the importance of the model violation in the metric of the observed score scale. So the effect sizes are on a scale ranging from 0 to the maximum item score. As a rule of thumb, effect sizes of more than 0.10 can be considered indicative of more than minor model violations with dichotomous items where the maximum item score is 1 [23]. The rule can be applied to polytomous items by multiplying the maximum item score by 0.1 and using that as an indication of more than minor model violations. The effect sizes are considered more important than the significant probabilities in this study because the power of the LM test increases with sample size and the sample sizes are very large in this study (14,076 in the Norwegian and 17,365 in the Scottish sample).

Once items that exhibit DIF have been identified, the next step is to model the DIF in such a way that the measures obtained at different time points are still comparable. This can be done by dividing each DIF item into several virtual items, one for each time point. Each virtual item is then given time point-specific item parameters, separating, e.g., holiday in 2002 and holiday in 2006 and 2010 statistically, even though they conceptually are the same item. In this way, it is assumed that the same construct is measured at all time points, but the item parameters may be different. Then, new item and person trait parameters are estimated and the fit of the data to the resulting model is assessed. If it can be shown that the items without DIF and the items with the time point-specific item parameters fit a concurrent Rasch model, the conclusion that all items relate to the same underlying construct is supported [25, 30].

The item parameters for the FAS were estimated by marginal maximum likelihood (MML) [31]. DIF and fit within subpopulations were examined using the LM statistic. These were calculated using the free Multidimensional Item Response Theory (MIRT) software package [32]. Additional analyses to check for the assumptions of unidimensionality [33] and local independence were conducted using the RUMM2030 software package [34].

Results

The first analysis was conducted to test whether the data from the FAS fit the Rasch model across the three survey years (2002, 2006 and 2010) for the Norwegian and Scottish samples, respectively. The results of the unidimensionality test [33] failed to reject the unidimensionality assumption (0.31 and 0.55 % of the *t* tests were significant in the Norwegian and Scottish data, respectively, which is

well below the nominal level of 5 %). Furthermore, a test of local independence showed that there were no items with positive residuals over 0.2 in the residual correlation matrix indicating acceptable fit to the Rasch model. However, there was evidence of significant DIF across survey years indicating IPD for the holiday and computer items in both the Norwegian and the Scottish samples (Table 1). There was no DIF for the car or bedroom items. Since there was DIF across survey years, it was important to establish fit to the Rasch model separately for each survey year before modeling the DIF across survey years. The effect sizes for all four items were below critical values on the LM test for all three survey years in both samples (Table 1).

An analysis was conducted in which the item parameters were calculated independently at each survey year. These results are reported in Table 2.

The mean threshold parameters at each survey year are also illustrated in Fig. 1 for the four items in the FAS, with the aim of graphically depicting the IPD across time. There is a clear trend that the average item parameter threshold in the computer item decreases across the survey years in both samples, indicating that the value of having a certain number of computers decreases over the survey years. However, Fig. 1 also illustrates that the IPD was not as clear for the holiday item which also exhibited DIF across survey years. Finally, the car and bedroom items did not display high levels of IPD across survey years which was expected based on the lack of DIF across time.

Since DIF was identified in the holiday and computer items, it was considered necessary to model this DIF by splitting these items into virtual items with survey year-specific item parameters. This process was done iteratively by splitting the item that had the largest DIF first, which is the recommended process in the Rasch model because spurious DIF can often occur in non-DIF items as a consequence of items with real DIF [22, 35, 36]. The computer item was split into three virtual items with time point-specific item parameters, and the DIF of the remaining three items was investigated. The effect size values <0.10 showed that there was no DIF in the remaining items in the Norwegian or the Scottish samples, respectively. Consequently, a single-item parameter was used for the car, bedroom and holiday items; and survey year-specific item parameters were used in estimating trait parameters for the computer item. Concurrent analyses using the LM test were conducted to test whether the combined model including the items with group-specific item parameters fit the data. Table 3 shows that all items including the items with group-specific item parameters fit the Rasch model.

With the purpose of illustrating the implications of ignoring IPD across survey years, analyses were performed to compare the mean scores before and after equating the

Table 1 Lagrange multiplier tests of DIF and overall fit for items in original analysis

Item	DIF	Norway			DIF	Scotland		
		Item fit for each survey year				Item fit for each survey year		
		2002	2006	2010		2002	2006	2010
Car	0.01	0.01	0.02	0.01	0.04	0.03	0.05	0.06
Bedroom	0.01	0.00	0.01	0.01	0.01	0.01	0.00	0.00
Holiday	0.20*	0.05	0.06	0.08	0.12*	0.03	0.04	0.04
Computers	0.21*	0.02	0.03	0.01	0.17*	0.01	0.02	0.03

The values reported are the LM effect sizes

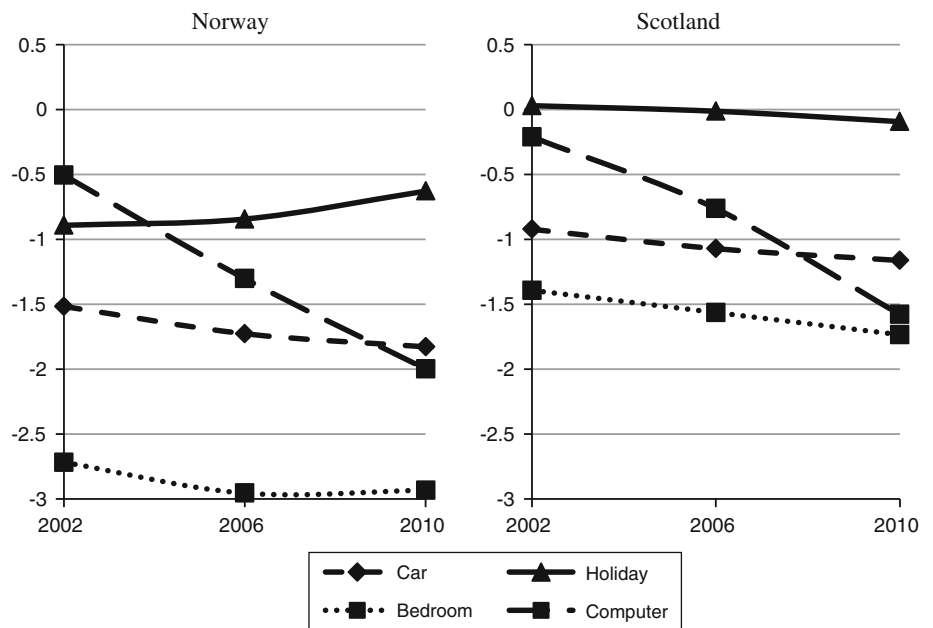
* Indicates DIF effect sizes over 0.10

Table 2 Item parameter estimates by year for the four FAS items in Norway and Scotland

Country	Item	2002			2006			2010		
		Thr. 1	Thr. 2	Thr. 3	Thr. 1	Thr. 2	Thr. 3	Thr. 1	Thr. 2	Thr. 3
Norway	Car	-2.82	-0.22		-3.02	-0.44		-2.82	-0.84	
	Bedroom	-2.72			-2.96			-2.93		
	Holiday	-1.76	-0.28	-0.63	-1.72	-0.21	-0.61	-1.34	-0.09	-0.46
	Computers	-2.52	0.27	0.72	-3.34	-0.44	-0.08	-3.25	-1.42	-1.31
Scotland	Car	-1.76	-0.09		-1.86	-0.28		-1.86	-0.47	
	Bedroom	-1.39			-1.56			-1.74		
	Holiday	-1.07	0.44	0.72	-1.06	0.31	0.71	-0.95	0.18	0.49
	Computers	-1.95	0.68	0.64	-2.93	0.04	0.60	-3.44	-0.96	-0.34

Thr threshold

Fig. 1 Graphical depiction of item parameter drift across survey years for the four FAS items. Note: The values shown are the mean threshold parameters at each survey year on the logit scale



scale across survey years. These results are presented for the Norwegian followed by the Scottish samples.

In the Norwegian sample, the mean and standard deviation (SD) for the analysis where IPD was ignored (prior to splitting the item with DIF) in 2002 was 6.25 (SD = 1.61), which increased to 6.75 (SD = 1.61) in 2006 and 7.22 (SD = 1.52) in 2010 (Table 4). Accounting for DIF in the model had an effect on the mean scores across the survey

years (Table 4). The mean of the 2002 remained the same at 6.25 (SD = 1.61) because this survey year was used as the reference. However, the mean for the 2006 survey sample decreased by 0.48–6.27 (SD = 1.56), and the mean for the 2010 sample decreased by 1.11–6.11 (SD = 1.53) after splitting the DIF item into virtual items with survey year-specific item parameters. Therefore, accounting for DIF due to IPD across survey years shrunk the average

Table 3 Lagrange multiplier tests of DIF and overall fit for items after splitting the DIF item into virtual items with time point-specific item parameters

Item	DIF	Norway			DIF	Scotland		
		Item fit for each survey year				Item fit for each survey year		
		2002	2006	2010		2002	2006	2010
Car	0.05	0.06	0.02	0.09	0.01	0.03	0.06	0.01
Bedroom	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.01
Holiday	0.06	0.08	0.05	0.10	0.02	0.04	0.04	0.05
Computers		0.03	0.03	0.01		0.00	0.01	0.03

The values reported are the LM effect sizes

Numbers in bold represent the item that is split into virtual items with group-specific item parameters

growth in apparent family affluence across time in the Norwegian sample from 0.50 to 0.02 from 2002 to 2006; from 0.47 to -0.16 from 2006 to 2010; and from 0.97 to -0.14 from 2002 to 2010. Figure 2 is a graphical depiction of the results. This means that the bias in the FAS scores in

2010 due to IPD was 1.11 raw score units or 0.73 SD units when ignoring the lack of MI across survey years in the Norwegian data.

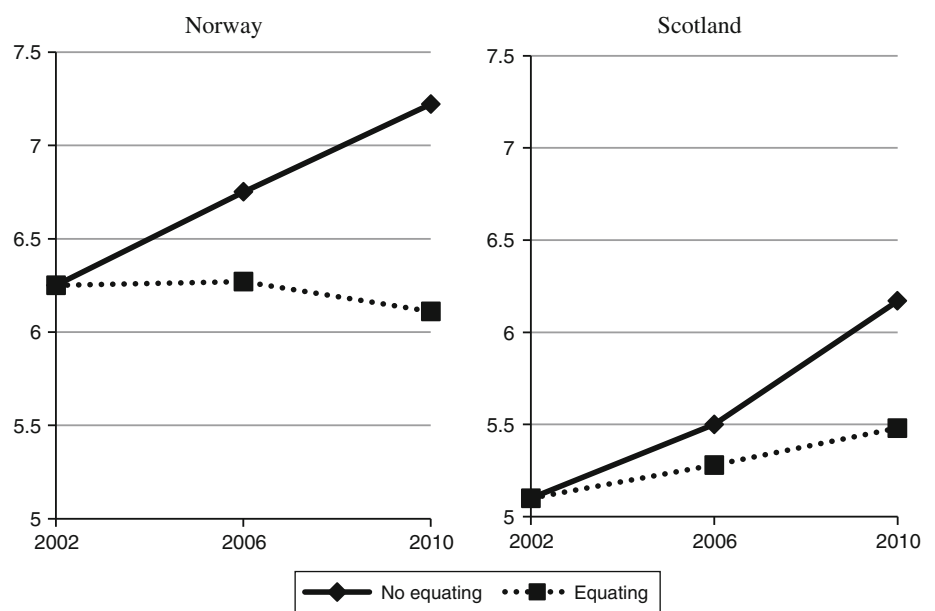
In the Scottish sample, the mean and SD for the analysis where IPD was ignored (prior to splitting item with DIF) in 2002 was 5.10 (SD = 1.85), this value increased to 5.50 (SD = 1.86) in 2006, and 6.17 (SD = 1.79) in 2010. As seen in the Norwegian data, accounting for DIF in the model had an effect on the mean scores across the survey years (Table 4). The mean of the 2002 remained the same at 5.10 (SD = 1.85) because this survey year was used as the reference. However, the mean for the 2006 survey sample decreased by 0.22–5.28 (SD = 1.93), and the mean for the 2010 sample decreased by 0.69–5.48 (SD = 1.90) after splitting the DIF item into virtual items with survey year-specific item parameters. Consequently, accounting for DIF due to IPD across survey years shrunk the average growth in apparent family affluence across time in the Scottish sample from 0.40 to 0.18 from 2002 to 2006; from 0.67 to 0.20 from 2006 to 2010; and from 1.07 to 0.38 from 2002 to 2010. Figure 2 is a graphical depiction of the results. This means

Table 4 Mean and SD of FAS scores in each survey year

Survey year	N	Norway				N	Scotland			
		Ignoring item parameter drift		Equating with time point-specific item parameters			Ignoring item parameter drift		Equating with time point-specific item parameters	
		Mean	SD	Mean	SD		Mean	SD	Mean	SD
2002	4,949	6.25	1.61	6.25	1.61	4,233	5.10	1.85	5.10	1.85
2006	4,576	6.75	1.61	6.27	1.56	5,895	5.50	1.86	5.28	1.93
2010	4,270	7.22	1.52	6.11	1.53	6,554	6.17	1.79	5.48	1.90

Only respondents with complete response patterns were used in this analysis

Fig. 2 Mean FAS scores when ignoring MI across survey years (in bold), and after accounting for MI by equating FAS scores across survey years with time point-specific item parameters (spotted line)



that the bias in the FAS scores in 2010 due to IPD was 0.69 raw score units or 0.39 SD units when ignoring the lack of MI across survey years in the Scottish data.

Discussion

The FAS has been used widely to investigate associations between SEP and various health outcomes and in trend analyses [7–9]. One of the greatest challenges to using the FAS in this type of research is that violations of MI have consistently been found across different samples, countries and across survey years [19, 20, 37, 38]. Inaccurate inferences can be made if MI cannot be established across these conditions [17]. Several studies have illustrated that DIF and MI are the rule rather than the exception in health-related scales including the FAS [19, 20]. This is specifically the case for the FAS because general societal trends that are not directly related to family affluence have a direct impact on the items, and because the scale is only made up of four items so IPD in one item can have great consequences for the conclusions that are made with the scale. When the comparison of family affluence is necessary over different HBSC survey years or when the longitudinal implications of family affluence are of interest, it is necessary to account for these trends in a systematic way in order to ensure that the measurement of the family affluence construct is not biased by these unrelated trends.

The results of the present study indicated that the FAS captures economic growth, but the scores are upwardly biased due to IPD across survey years. The effect size of the bias was 1.11 raw score units (0.73 SD units) in the Norwegian and 0.69 raw score units (0.39 SD units) in the Scottish sample in 2010. This means that inaccurate inferences will often be made when the FAS scores are obtained by adding the scores on each item to produce a sum score, and interpretations are made based on changes in these sum scores. More precisely, using the FAS in its current format without accounting for IPD across survey years would result in the conclusion that the FAS is increasing quite consistently in Norway and Scotland. However, the results show that a large part of the increase in the FAS scores can be attributed to bias in the FAS because of the use of the computer item which exhibited IPD across time. The increase in the FAS was more modest in Scotland and slightly negative in Norway once the DIF in the computer item was accounted for in this study. The effect sizes of bias reported in this study may have large implications for the conclusions that are made when the FAS is used as a primary variable such as in the assessment of children's eligibility for free school meals [5], or when the FAS is used as an independent variable to investigate associations between SEP and various health outcomes [8, 11].

In the original analysis for the Norwegian and Scottish samples, both the computer and holiday item exhibited DIF across survey years. However, the DIF in the holiday item disappeared once the DIF in the computer item was accounted for. This is a typical finding in DIF research when using the Rasch model [35]. Generally, if there is real DIF in some items which favor one group, then as an artifact of this procedure, artificial DIF that favors the other group is induced in the other items. In the present study, an iterative procedure for detecting items with real DIF with the purpose of identifying items that may have no DIF was used according to suggestions from the literature [35]. Consequently, the results of the study indicate that there was no systematic DIF in the remaining three items across survey years for the Norwegian or the Scottish data.

Implications for policy and practice

The results of this study have pointed to a need to account for IPD in interpretation of changes in family affluence across time. Since DIF was only found in one the FAS item in this study, an argument could be made that the computer item should be eliminated from the scale. This argument may be problematic as the item conceptually is an important indicator of family affluence, so a solution is to accurately model the IPD to increase validity of the scale. Although other items did not exhibit DIF across time once the computer item was eliminated, there is a possibility that trends such as changes in airfare prices or large increases in the prices of gas could induce DIF in holiday and car items, respectively. There may also be trends in specific countries that could affect other the FAS items; therefore, the method described in this study seems vital for providing a means of assessing and accounting for IPD in the use of the FAS.

The equating method using the Rasch model described in this study could essentially be used in increasing the number of items in the FAS or in increasing the number of response categories in the existing items. An example where this could be appropriate is the computer item. In the current version of the FAS, the response options to the question “How many computers does your family own?” include: None = 0, One = 1, Two = 2, More than two = 3. In 2002, 19 % of the Norwegian and 17 % of the Scottish sample stated that they had more than two computers. These percentages increased to 40 and 24 % in 2006; and 72 and 48 % in 2010 for the Norwegian and Scottish samples, respectively. Therefore, the “more than two” option may no longer be appropriate for differentiating between adolescents with high and low FAS in future HBSC surveys. One option could be to include more categories in order to accurately differentiate across the sample. Another option could be to introduce new items

and assess if there is MI across the two versions of the FAS.

Another possible explanation for the IPD in the computer item is that the meaning of the item might have changed over time. Today, asking people about how many computers they have is an ambiguous question. It is not clear whether this includes computer tablets and/or smartphones. Therefore, it is possible that the interpretation of the item has changed which means that the item needs to be reworded.

With the current knowledge on validity of the FAS produced from research with a conceptual or theoretical [38] as well as a statistical focus [19, 20, 37], and the results from the current study, it is clear that when researchers include the FAS in comparative studies across time, they should be aware of IPD when interpreting their results.

Limitations and future research

Although the sample sizes were very large (a total of 31,441), which is a strength in the present study. One limitation in the study is the relative homogenous data samples from two European countries (Norway and Scotland), meaning that the results cannot be generalized to all HBSC countries where different trends may introduce alternative patterns of IPD in the FAS items. Future research could benefit from applying the method described in the present study to other HBSC countries and regions.

Another limitation is the use of a method for detecting and accounting for IPD with a scale consisting of as few as four items, meaning that the potential effects of DIF on one item are quite large for the FAS. Although this is not specifically a limitation to this study, but rather a limitation in the FAS because it only includes four items, it is important to note that methodological criteria for identifying DIF have a large impact on the results. Although there is no empirical evidence for the stability of the LM test with 4 items, simulation studies have shown that the LM test procedure used in this study has good power and expected Type I error values in tests of 10 items and sample sizes over 1,000 when there are 30 % or fewer DIF items in the test [28, 29]. The results of these studies show that the stability of the statistic in terms of power and Type I error depends on the number of items in the test, sample size and number of DIF items. Here, the stability of the statistic goes up with more items, larger sample sizes and fewer DIF items. Based on these results, we expect that the test would function well here as the sample sizes were 14,076 and 17,365 for the Norwegian and Scottish samples, respectively. Nonetheless, we would suggest cautious use of the statistic for the FAS with small sample sizes.

Similarly, effect sizes were used instead of significance probabilities in this study because even smallest deviations from the model became significant with large sample sizes such as those used in this study. Rules of thumb based on previous research [23] were used here; however, there are many methods for identifying DIF, and future research could investigate the use of different methods because a more conservative method could have a large impact on the results if two rather than one item is identified with true DIF across time points.

Future research could also be used to assess MI and account for IPD in future HBSC surveys where new trends may introduce additional IPD in the current the FAS items. Finally, future research should also investigate the consequences of using the equating method described in this study by assessing how the FAS relates to important outcomes and external variables when the DIF-equated solution for scoring the FAS is used.

References

1. Mackenbach, J. P. (2002). Evidence favouring a negative correlation between income inequality and life expectancy has disappeared. *British Medical Journal*, *324*(7328), 1–2.
2. Carstairs, V., & Morris, R. (1989). Deprivation: Explaining differences in mortality between Scotland and England and Wales. *British Medical Journal*, *299*, 886–889.
3. Currie, C., 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*(3), 385–397.
4. Wardle, J., Robb, K., & Johnson, F. (2002). Assessing socioeconomic status in adolescents: The validity of a home affluence scale. *Journal of Epidemiology Community Health*, *56*(8), 595–599.
5. Kehoe, S., & O'Hare, L. (2010). The reliability and validity of the Family Affluence Scale. *Effective Education*, *2*(2), 155–164.
6. Currie, C., Molcho, M., Boyce, W., Holstein, B., Torsheim, T., & Richter, M. (2008). Researching health inequalities in adolescents: The development of the Health Behaviour in School-Aged Children (HBSC) Family Affluence Scale. *Social Science and Medicine*, *66*(6), 1429–1436.
7. Boyce, W., Torsheim, T., Currie, C., & Zambon, A. (2006). The Family Affluence Scale as a measure of national wealth: Validation of an adolescent self-report measure. *Social Indicators Research*, *78*(3), 473–487.
8. Torsheim, T., Ravens-Sieberer, U., Hetland, J., Valimaa, R., Danielson, M., & Overpeck, M. (2006). Cross-national variation of gender differences in adolescent subjective health in Europe and North America. *Social Science and Medicine*, *62*(4), 815–827.
9. Zambon, A., Boyce, W., Cois, E., Currie, C., Lemma, P., Dalmasso, P., et al. (2006). Do welfare regimes mediate the effect of socioeconomic position on health in adolescence? A cross-national comparison in Europe, North America, and Israel. *International Journal of Health Services*, *36*(2), 309–329.
10. Andersen, A., Krolner, R., Currie, C., Dallago, L., Due, P., Richter, M., et al. (2008). High agreement on family affluence

- between children's and parents' reports: International study of 11-year-old children. *Journal Epidemiology Community Health*, 62, 1092–1094.
11. 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), 1–12.
 12. Townsend, P. (1987). Deprivation. *Journal of Social Policy*, 16, 125–146.
 13. Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.
 14. Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. London: Lawrence Erlbaum Associates.
 15. Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement*, 22(1), 33–51.
 16. van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing*. New York: Springer.
 17. Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3, 4–69.
 18. Makransky, G., & Glas, C. A. W. (2013). Modeling differential item functioning with group-specific item parameters: A computerized adaptive testing application. *Measurement*. doi:10.1016/j.measurement.2013.06.020.
 19. Schnohr, C. W., Kreiner, S., Due, E., Currie, C., Boyce, W., & Diderichsen, F. (2008). Differential item functioning of a Family Affluence Scale: Validation study on data from HBSC 2001/02. *Social Indicators Research*, 89(1), 79–95.
 20. Schnohr, C. W., Makransky, G., Kreiner, S., De Clercq, B., Hofmann, F., Torsheim, T., Elgar, F., & Currie, C. (2013). Item response drift in the Family Affluence Scale: A study on three consecutive surveys of the Health Behavior in School-aged Children (HBSC) survey. *Measurement*, 45(9), 3119–3126.
 21. Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis and Rheumatism*, 57(8), 1358–1362.
 22. Kreiner, S., & Christensen, K. B. (2011). Item screening in graphical loglinear Rasch models. *Psychometrika*, 76(2), 228–256.
 23. van Groen, M. M., ten Klooster, P. M., Taal, E., van de Laar, M. A. F. J., & Glas, C. A. W. (2010). Application of the health assessment questionnaire disability index to various rheumatic diseases. *Quality of Life Research*, 12, 1255–1263.
 24. Weisscer, N., Glas, C. A., Vermeulen, M., & De Haan, R. J. (2010). The use of an item response theory-based disability item bank across diseases: Accounting for differential item functioning. *Journal of Clinical Epidemiology*, 63, 543–549.
 25. Glas, C. A. W. (1999). Modification indices for the 2-pl and the nominal response model. *Psychometrika*, 64, 273–294.
 26. Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174.
 27. Glas, C. A. W. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica*, 8, 647–667.
 28. Kalid, M. N. (2009). IRT model fit from different perspectives. Doctoral dissertation, University of Twente, The Netherlands.
 29. Kalid, M. N., & Glas, C. A. W. (2013). A scale purification procedure for evaluation of differential item functioning. Manuscript submitted for publication.
 30. Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models. Their foundations, recent developments and applications* (pp. 69–96). New York: Springer.
 31. Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443–459.
 32. Glas, C. A. W. (2010). *MIRT: Multidimensional Item Response Theory. (Computer Software)*. University of Twente. <http://www.utwente.nl/gw/omd/afdeling/Glas/>.
 33. Smith, E. V., Jr. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, 3(2), 205–231.
 34. Andrich, D., Sheridan, B., & Luo, G. (2010). *Rasch models for measurement: RUMM2030*. Perth: RUMM Laboratory.
 35. Andrich, D., & Hagquist, C. (2012). Real and artificial differential item functioning. *Journal of Educational and Behavioral Statistics*, 37(3), 387–416.
 36. Christensen, C. B., Kreiner, S., & Mesbah, M. (2013). *Rasch model in health*. New York: Wiley.
 37. Batista-Foguet, J. M., Fortinana, J., Currie, C., & Villalbi, J. R. (2004). Socio-economic indexes in surveys for comparisons between countries. *Social Indicators Research*, 67, 315–332.
 38. Elgar, F. J., De Clercq, B., Schnohr, C. W., Bird, P., Pickett, K. E., Torsheim, T., et al. (2013). Absolute and relative family affluence and psychosomatic symptoms in adolescents. *Social Science and Medicine*, 91, 25–31.