JOURNAL OF APPLIED MEASUREMENT, *17*(3), 312-334

Differential Item Functioning (DIF) and Subsequent Bias in Group Comparisons using a Composite Measurement Scale: A Simulation Study

Alexandra Rouquette
AP-HP, Hôtel-Dieu Hospital, Biostatistics and Epidemiology Department, Paris, France
Inserm, U1178, Mental Health and Public Health,
Université Paris-Sud and Université Paris Descartes, Paris, France
EA 4360 - Research unit APEMAC, Nancy-Université, Université Paris-Descartes,
Université Metz Paul Verlaine, Paris, France

Jean-Benoit Hardouin
EA 4275 – SPHERE (Biostatistics, Pharmacoepidemiology
and Human Sciences Research), University of Nantes, France
Biostatistics and Methodology Unit, CHU de Nantes, Nantes, France

Joël Coste
AP-HP, Hôtel-Dieu Hospital, Biostatistics and Epidemiology Department, Paris, France
EA 4360 - Research unit APEMAC, Nancy-Université, Université Paris-Descartes,
Université Metz Paul Verlaine, Paris, France

Objective. To determine the conditions in which the estimation of a difference between groups for a construct evaluated using a composite measurement scale is biased if the presence of Differential Item Functioning (DIF) is not taken into account.

Methods. Datasets were generated using the Partial Credit Model to simulate 642 realistic scenarios. The effect of seven factors on the bias on the estimated difference between groups was evaluated using ANOVA: sample size, true difference between groups, number of items in the scale, proportion of items showing DIF, DIF-size for these items, position of these items location parameters along the latent trait, and uniform/non-uniform DIF.

Results. For uniform DIF, only the DIF-size and the proportion of items showing DIF (and their interaction term) had meaningful effects. The effect of non-uniform DIF was negligible.

Conclusion. The measurement bias resulting from DIF was quantified in various realistic conditions of composite measurement scale use.

Requests for reprints should be sent to Alexandra Rouquette, Hôtel-Dieu Hospital, Biostatistics and Epidemiology Department, 1 place du parvis Notre-Dame, 75181 Paris cedex 04, France, e-mail: alex.rouquette@gmail.com.

Introduction

Composite measurement scales are increasingly used in health research, particularly for work on patients' self-perception of health outcomes including, for example, quality of life, satisfaction with care, anxiety or disability. Most health research questions involve comparisons between groups. Is the level of health-related quality of life the same in two countries? Are the effects of stressful events on the level of anxiety the same in men and women? Is the satisfaction with care related to age? To answer such questions correctly, all sources of bias in the study have to be considered, and especially one which is of particular interest in this context: the measurement bias which can result from the presence of differential item functioning (DIF) in the scale.

DIF describes the phenomenon of one or several items of a questionnaire "functioning" differently in the groups of individuals to be compared (defined by a characteristic such as age, sex or country of origin for example) (Mellenbergh, 1989; Millsap and Everson, 1993). Statistically, it means that the parameters of the function relating the latent variable (the measured construct: anxiety, quality of life, etc.) to the observations (responses to the scale items) is different in the various groups involved in the comparison (Borsboom, 2006). For example, if the data are analyzed using a Rasch model, the item response function expresses the conditional probability of responding positively to a binary item, given the individual's status on the latent variable (or latent trait, θ), as a logistic function of θ and an item location parameter (δ) called the difficulty parameter (Rasch, 1960). A DIF phenomenon exists if the difficulty parameter differs between the groups compared (Meredith and Teresi, 2006; Millsap, 2011).

Various methods have been proposed to test for the presence of DIF in a scale; some use latent variable models, such as the Rasch model, and others use only the observed variables (item answers and the scale score) (Millsap, 2011). The most widely used in the literature are based on nested-model tests, the logistic regression methods or the Mantel-Haenszel test. Whatever the method used, DIF detection involves a statistical test in which the null hypothesis is "There is no DIF" (i.e., "The constrained and unconstrained models fit the data equally") and the alternative hypothesis is "There is DIF." Therefore, conclusions emerging from studies using these methods are strongly influenced by the sample size: indeed, with sufficiently large sample sizes, DIF would probably be detected in all the items of all scales (Borsboom, 2006). Thus, in practice, the important issue is the meaningfulness of the DIF found within the scale rather than its statistically significance.

Various indices of DIF effect-size at the item level have been published. If a Rasch model is used, a widely cited criterion for a meaningful DIF is a difference greater than 0.5 logit between the estimated item difficulty parameters in the two groups to be compared ($|\delta_{dif}|$) (Jarl, Heinemann, and Norling Hermansson, 2012; Kenaszchuk, Wild, Rush, and Urbanoski, 2013; Lai, Cella, Chang, Bode, and Heinemann, 2003; Linacre, 1994; Steinberg and Thissen, 2006; Tennant and Pallant, 2007; Tristan, 2006). The Educational Testing Service (ETS) DIF classification rule is another widely used criterion; it was developed primarily for the Mantel-Haenszel DIF detection method and later adapted for use on the same scale of the Rasch difficulty parameters (Dorans and Holland, 1992; Paek and Wilson, 2011). Three DIF-levels are defined in this classification: Negligible ($|\delta_{dif}| < 0.436$), Intermediate ($0.436 \leq |\delta_{dif}| < 0.638$) and Large ($|\delta_{dif}| \geq 0.638$). No explanation has been given concerning the thresholds used in both 0.5 logit and ETS rules. Other criteria have been defined for logistic regression methods to qualify the magnitude of DIF at the item level, but once again, without justification being made available (Barbier, Peters, and Hansez, 2009; Bjorner and Pejtersen, 2010; Crane, Hart, Gibbons, and Cook, 2006; Zumbo, 1999).

Due to the absence of a clear comprehension of the practical meaning of a statistically significant DIF, a trend is emerging in empirical studies aimed at determining the presence of items showing DIF (DIF-items) in a composite measurement scale: when statistically significant DIF-items are found, their effect on the conclusion drawn at the scale level is also evaluated. This has mainly been done in work comparing the results obtained using the entire scale to those obtained using the scale without DIF-items (Bjorner, Kreiner, Ware, Damsgaard, and Bech, 1998; Czachowski, Terluin, Izdebski, and Izdebski, 2012; Earleywine, LaBrie, and Pedersen, 2008; Fleishman, Spector, and Altman, 2002; Goetz et al., 2011; King, Street, Gradus, Vogt, and Resick, 2013; Lange, Thalbourne, Houran, and Lester, 2002; Morales, Reise, and Hays, 2000), or with or without adjustment for the presence of DIF-items in the analyses (Banh et al., 2012; Coste et al., 2014; Crane et al., 2007, 2010, 2008; Gibbons et al., 2009; Hart, Deutscher, Crane, and Wang, 2009; Jones and Gallo, 2002; Jones, 2003; Orlando and Marshall, 2002; Petersen et al., 2010; Rodriguez and Crane, 2011; Song, Cai, Brown, and Grimm, 2011; Wanders et al., 2015; Woodbury et al., 2008; Yu, Yu, and Ahn, 2007). Reversal of the conclusion of the study has rarely been clearly observed (Fleishman et al., 2002; Song et al., 2011; Yu et al., 2007); however, DIF-items have in some cases been found to be responsible for a part of the difference detected between the groups compared (Crane et al., 2007, 2010; Jones and Gallo, 2002; Jones, 2003; Lange et al., 2002). Nevertheless, in most of these studies, the presence of DIF-items was found to have a “negligible” to “small” effect at the scale level (Banh et al., 2012; Barbier et al., 2009; Bjorner et al., 1998; Crane et al., 2007, 2006, 2008; Czachowski et al., 2012; Earleywine et al., 2008; Gibbons et al., 2009; Hart et al., 2009; Jarl et al., 2012; Kenaszchuk et al., 2013; King et al., 2013; Morales et al., 2000; Orlando and Marshall, 2002; Petersen et al., 2010; Rodriguez and Crane, 2011; Wanders et al., 2015; Woodbury et al., 2008).

As yet, there is no clear recommendation on what to do if DIF-items are found in a questionnaire. Removing them from the scale is not without consequences on the psychometric properties; however, if various characteristics (sex, age, ethnicity, etc.) are affected by a DIF phenomenon within a scale, it may be difficult to adjust for all sources of DIF in multivariate analyses (Meade, 2010). It would therefore be extremely valuable to clarify the conditions in which DIF-items have a practical meaningful effect at the scale level (Borsboom, 2006; Tennant and Pallant, 2007). Simulation studies are a useful approach to evaluate DIF effects on a chosen inference (prevalence estimates, difference between groups, etc.) in various controlled conditions (Teresi, Ramirez, Jones, Choi, and Crane, 2012). Three simulation studies, using item response theory models, have been performed and have given information on factors (magnitude of the DIF phenomenon at the item level, number of DIF-items within the scale, sample and group sizes, magnitude of the mean latent trait level difference between groups) which could influence the size of the DIF-associated bias affecting the observed mean score difference between groups (Golia, 2010; Lee and Zhang, 2010; Li and Zumbo, 2009). However, the conditions used in these simulation studies (number of items in the scale larger than 15 and sample sizes larger than 500) were not representative of those encountered in health research studies in which the number of items in the questionnaire and the sample sizes are often small, and the use of Rasch models is often preferred to detect DIF phenomenon (Christensen, Kreiner, and Mesbah, 2012; Hardouin et al., 2012; Sébille, Blanchin, Guillemin, Falissard, and Hardouin, 2014).

The aim of this simulation study was to determine the conditions, amongst those encountered in health research studies, in which the estimation of a difference between two groups for a construct evaluated using a questionnaire is biased if the presence of DIF-items concerning the group under study is not taken into account. A model for polytomous data belonging to the Rasch measurement theory, the partial credit model (PCM), was chosen to simulate and analyze data because polytomous items are frequently used in health research studies, and also because it allows the effects of two kinds of DIF to be studied. Uniform DIF is defined as holding when the difference in the location parameters of a DIF-item between the two groups is the same at all levels of the latent variable, otherwise, the DIF is called nonuniform

(Mellenbergh, 1989). To be able to disentangle the effect of each DIF-related factor studied (p = proportion of the items in the scale that are DIF-items; δ_{dif} = DIF size for each DIF-item; *pos* = position of the DIF-items location parameters along the latent trait and kind of DIF), a representative scenario of health research studies was simulated, in which the different factors of interest could be manipulated: the determination of the difference of the mean level of a construct measured on the latent trait scale (γ) using a composite measurement scale (J items with five response categories each) between two groups of equal size (N).

Methods

Data generation

Data were generated using the PCM in which the probability of a response y to an item j (j = 1,…,J) with K_j + 1categories (k = 0,…,K_j) for the subject i (i = 1,…,N) is a function of the subject's latent trait level (θ_i) (Masters, 1982). This model can be written:

$$P\left(Y_{ij}=ky\mid\theta_i,\delta_{jk}\right)=\frac{\exp\left(y\theta_i-\sum_{k=1}^{y}\delta_{jk}\right)}{\sum_{c=0}^{K_j}\exp\left(c\theta_i-\sum_{k=1}^{c}\delta_{jk}\right)},$$

where δ_{jk} is the item location parameter associated with the response category k of the item j and $\theta \sim N(\mu,1)$. The number of response categories for each item was set at five (K_j = 4 for j = 1,…, J). Two numbers of items in the scale (J = {4, 8}) and two group sizes (N = {100, 200}) were studied, the two groups to be compared being of equal size. It was decided to evaluate the influence of the magnitude of the difference in mean latent trait level between the two groups using three values, postulating that 0.1 standard deviation (*SD*) would be the hypothetical meaningful difference for the simulated scale (γ = {0, 0.1, 0.2}, i.e., μ was set at $\frac{-\gamma}{2}$ in the reference group and at $\frac{\gamma}{2}$ in the focal group). The values for the response category location parameters, δ_{jk}, were chosen as percentiles of the normal distribution; the example of the δ_{jk} values in the case of the eight-item scale is shown in Figure 1.

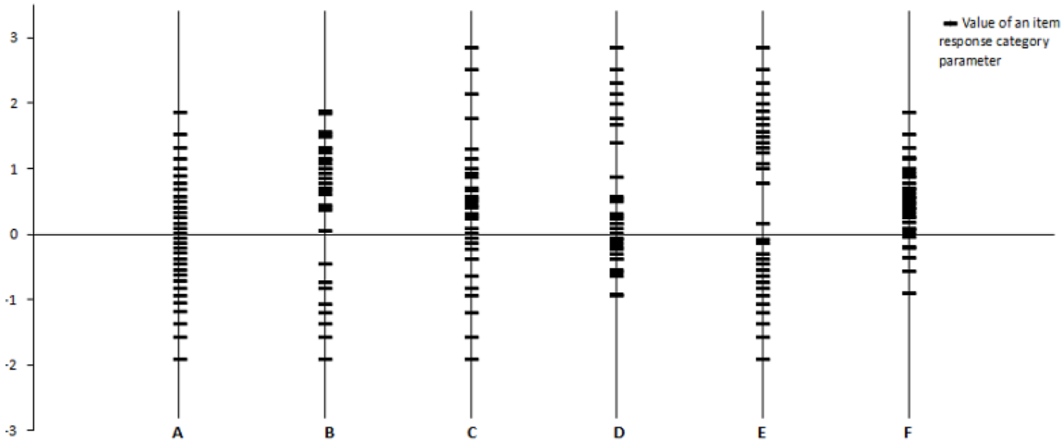
Uniform DIF

The bias resulting from the presence of DIF in the scale was first evaluated in the case of uniform DIF. Three DIF-related factors were studied: the proportion of items that were DIF-items in the scale (p = {0.25, 0.5, 0.75}), the DIF size for each DIF-item (δ_{dif}), and the position of the DIF-items location parameters along the latent trait continuum (*pos* = {Unif, Mean, Extreme, High, Low}); "Unif" meaning "DIF-items location parameters uniformly distributed along the latent trait continuum," "Mean" meaning "DIF-items location parameters close to the mean of the item difficulties in the scale," "Extreme" meaning "DIF-items are the most difficult and the easiest items of the scale," "High" meaning "DIF-items are the most difficult items in the scale," and finally "Low" meaning "DIF-items are the easiest items in the scale"). Figure 1 shows a graph of the values of each item response category location parameter along the latent trait continuum for the reference group (i.e., without DIF) and for each of the five positions of the DIF-items (i.e., "*pos*") in the focal group, in the case of a eight-item scale with 50% of DIF-items and a DIF-size, δ_{dif}, set at 1. Five hundred datasets were simulated for each of the 540 combinations of N, γ, J, p, δ_{dif} and *pos*, resulting in 270,000 datasets simulated in the case of uniform DIF.

Non-uniform DIF

Once the results concerning uniform DIF were known, non-uniform DIF-related bias was studied using only factors found to have a meaningful influence in the case of uniform DIF. The number of items in the scale J was set at 8, the group size N at 200 and the mean latent trait level difference between the two groups γ at 0.1. The same values as those studied in the case of uniform DIF were used for the three DIF-related factors: p = {0.25, 0.5, 0.75}, δ_{dif} = {0.25, 0.5, 1} and *pos* = {Unif, Mean, Extreme, High, Low}. A feature of the non-uniform DIF led to the study of two cases. In the first one, the sizes of DIF affecting the four DIF-item location parameters were:

$$-\frac{\delta_{dif}}{2},-\frac{\delta_{dif}}{4},+\frac{\delta_{dif}}{4},\text{and }+\frac{\delta_{dif}}{2},$$



	Value of the response category parameters				Items with DIF in the scale				
	A				B	C	D	E	F
	Without DIF								
					Around the mean	Uniformly distributed	Around the extremes	High difficulties	Low difficulties
Response category	1	2	3	4					
Item 1	−1.88	−1.55	−1.17	−0.80			X		X
Item 2	−1.34	−1.03	−0.70	−0.43		X	X		X
Item 3	−0.91	−0.60	−0.35	−0.11	X				X
Item 4	−0.52	−0.27	−0.04	0.19	X	X			X
Item 5	−0.19	0.04	0.27	0.52	X			X	
Item 6	0.11	0.35	0.60	0.91	X	X		X	
Item 7	0.43	0.70	1.03	1.34			X	X	
Item 8	0.80	1.17	1.55	1.88		X	X	X	

Figure 1. Values of the response category location parameters of the 8-item scale without differential item functioning (DIF) used to simulate data in the five cases concerning the position of the location parameters of items with DIF along the latent trait, when the size of DIF was set at 1 and the proportion of items in the scale which are items with DIF was set at 50% (θ: latent trait, X: item with DIF)

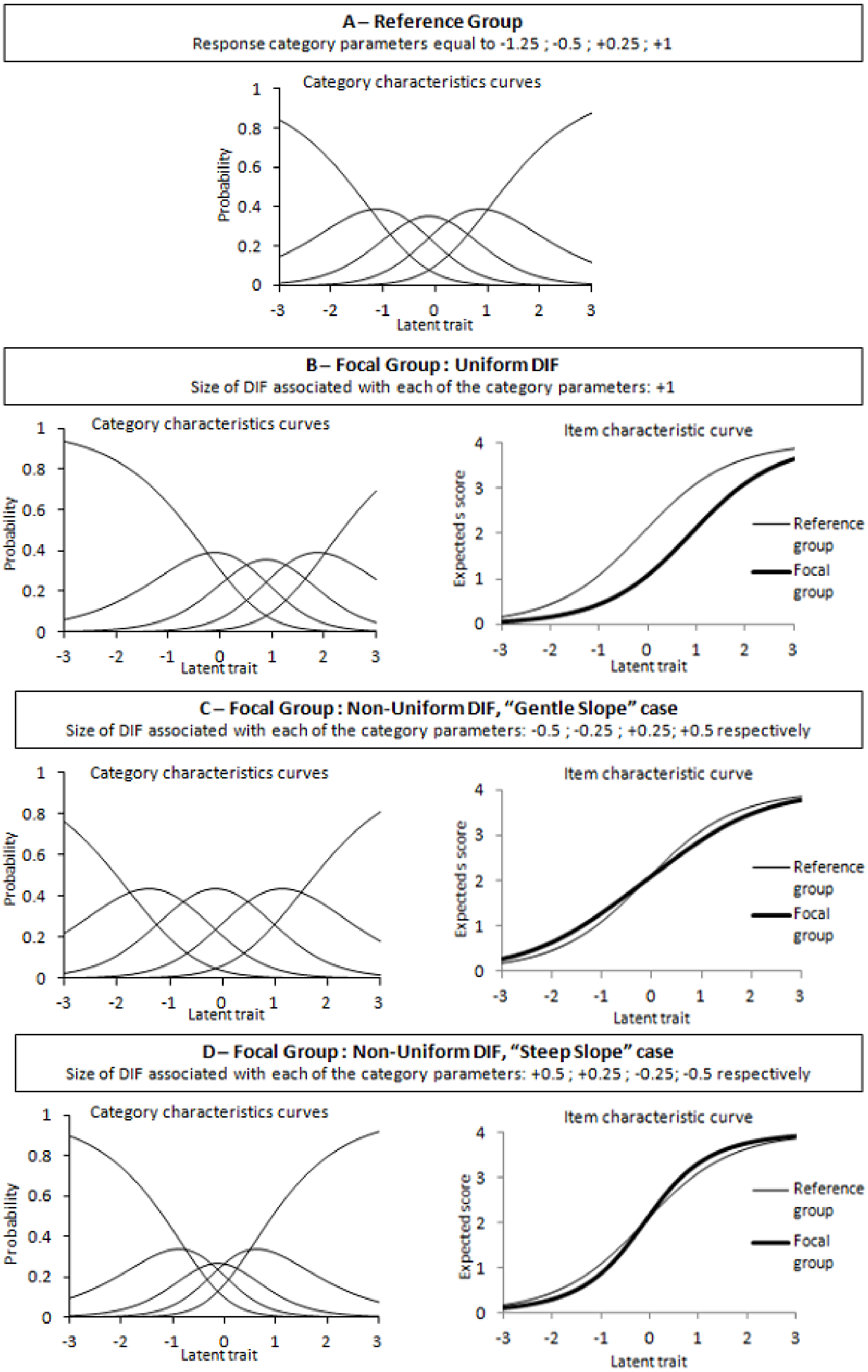
respectively. This resulted in a divergence of the four location parameters of the item along the latent trait continuum and, as a result, the slope of the item characteristic curve in the focal group was less steep than in the reference group (illustrated in Figure 2). This first case was called the "gentle slope" non-uniform DIF. The second case was called the "steep slope" non-uniform DIF, with the sizes of DIF affecting the four DIF-item location parameters being:

$$+\frac{\delta_{dif}}{2},+\frac{\delta_{dif}}{4},-\frac{\delta_{dif}}{4},\text{and }-\frac{\delta_{dif}}{2},$$

respectively. This resulted in a tightening of the four DIF-item location parameters along the latent trait continuum. A graph of the values of each item location parameter placed along the latent trait continuum is depicted in the Figure 3 (analogous to Figure 1), for each of the five positions of the DIF-items, in both cases of non-uniform DIF, in an eight-item scale with 50% of DIF-items and a DIF-size, δ_{dif}, set at 1. Five hundred datasets were simulated for each of the 90 combinations of p, δ_{dif}, *pos* and kinds of non-uniform DIF, resulting in 45,000 simulated datasets.

No DIF

To set a benchmark, five hundred datasets were also simulated for each of the 12 combinations of sample size N, mean latent trait level difference between the two groups γ and number of items in the scale J, with all the DIF-related factors set at zero.



A– Reference Group
Response category parameters equal to -1.25 ; -0.5 ; +0.25 ; +1

B– Focal Group : Uniform DIF
Size of DIF associated with each of the category parameters: +1

C– Focal Group : Non-Uniform DIF, "Gentle Slope" case
Size of DIF associated with each of the category parameters: -0.5 ; -0.25 ; +0.25; +0.5 respectively

D– Focal Group : Non-Uniform DIF, "Steep Slope" case
Size of DIF associated with each of the category parameters: +0.5 ; +0.25 ; -0.25; -0.5 respectively

Figure 2. Category and item characteristic curves for the same item affected by uniform or non-uniform ("gentle slope" or "steep slope") differential item functioning (DIF)

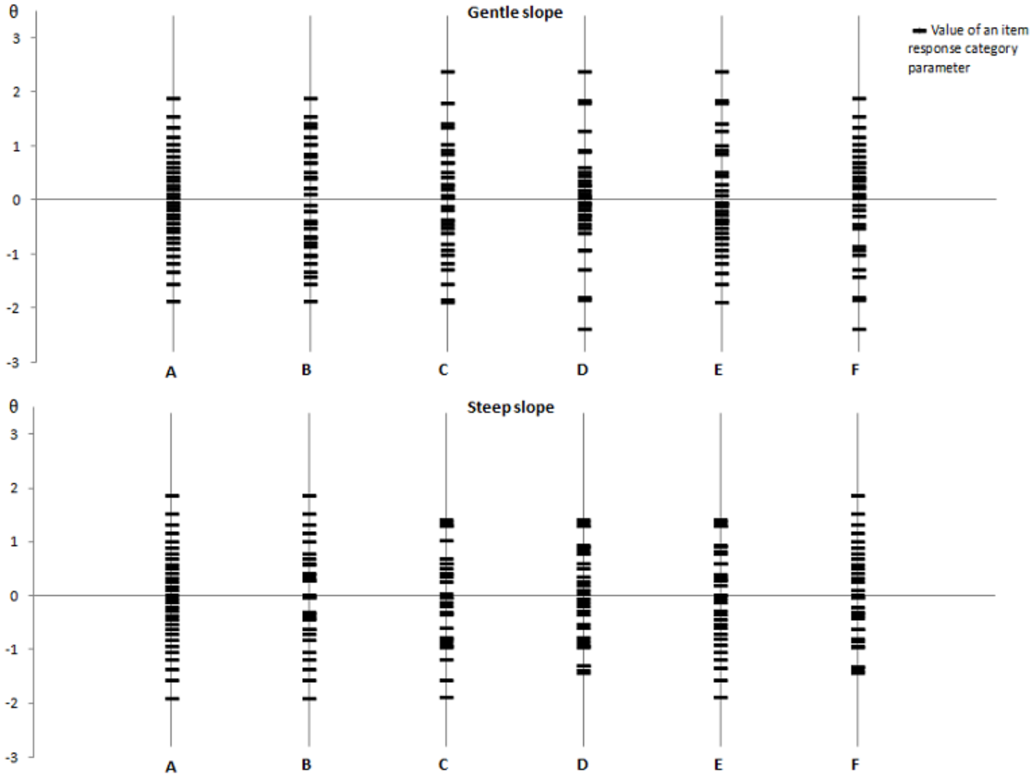


Figure 3. Values of the response category location parameters used to simulate data for a eight-item scale with 50% of items with DIF and a DIF-size set at 1, depending on the five cases of the position of the location parameters of items with DIF along the latent trait (A: without DIF, B: DIF-items uniformly distributed, C: DIF-items around the mean of the item difficulties, D: DIF-items with high and low difficulties, E: DIF-items with high difficulties, F: DIF-items with low difficulties, θ: latent trait)

Analyses of the simulated datasets

A latent regression PCM was applied to each of the simulated datasets to estimate the difference between groups (γ_{est}) without taking into account the presence of DIF in the scale. The DIF-related measurement bias was thus computed as bias = ($\gamma_{est} - \gamma$) and its mean (*SD*) was then described for each of the 540 and 90 factor combinations concerning the uniform and non-uniform DIF respectively. To obtain a complementary perspective, bias was dichotomized using the hypothetical meaningful difference (0.1) as the threshold. The frequency (%) of replications in which bias ≥ 0.1 was thus also determined for each factor combination. In the case of non-uniform DIF, in addition to bias ≥ 0.1, the frequency (%) of replications in which bias ≥ -0.1 was also evaluated because of the potentially two-directional influence of non-uniform DIF. For the datasets simulated without DIF, the random error was computed in the same way as bias (error = ($\gamma_{est} - \gamma$)) and its mean (*SD*) was determined for each of the 12 factor combinations; the frequency (%) of replications in which error ≥ 0.1 and that in which error ≥ -0.1 were also determined. Finally, the influence of each studied factor on bias was evaluated using a multivariate model of analysis of variance in both kinds of DIF (statistically significance if the *p*-value <0.05). Stata© software version 12 was used for data generation and statistical analyses (StataCorp, 2012).

Results

Simulations without DIF

The mean (*SD*) error and the frequency (%) of replications, among 500, in which it was higher than 0.1 or smaller than 0.1 are shown in Table 1, according to the sample size *N*, the mean latent trait level difference between the two groups γ and the number of items in the scale *J*. The mean random error was in all cases smaller than 0.013 and was not affected by the three factors manipulated. However, its *SD* was higher when the number of items *J* increased and was smaller when the sample size *N* increased. The same effects of the sample size and the number of items *J* were observed on the frequency (%) of replications in which the random error was higher than 0.1 or smaller than –0.1. These proportions were never higher than 35%.

Simulations with uniform DIF

The mean bias (*SD*) and the frequency (%) of replications among 500 in which when the DIF-size δ_{dif} was set to 0.25 are shown in Table 2A and 3A for a sample size set at 100 and 200 respectively, according to the other four factors manipulated. Tables for the DIF-size δ_{dif} set to 0.5 or 1 are shown in the appendix. In these descriptive analyses, no (or negligible) effect on the mean bias was observed for group size *N*, the number of items in the scale *J*, and the mean latent trait level difference between groups γ. However, as in the simulations without DIF (Table 1), the SD of the bias increased with the number of items *J* and decreased with increasing group size *N*. The effect was more notable for the three DIF-related factors (δ_{dif}, *p*, and, to a lower extent, *pos*). Table 4 reports the mean (*SD*) bias and the frequency (%) of replications in which bias $\geq$ 0.1 depending on the three DIF-related factors, with the group-size *N* set at 200, the number of items in the scale at 8 and and the mean latent trait level difference between groups γ at 0. The bias was higher when the DIF-size δ_{dif} and the proportion of DIF-items increased and, although the mean latent trait level difference between the two groups γ was set at 0, this bias exceeded the hypothetical meaningful difference in at least half of the replications for all the combinations of δ_{dif} and *p*, except when a DIF of 0.25 was simulated in only 25% items of the scale. The mean bias and the frequency (%) of replications in which bias $\geq$ 0.1 were however consistently higher than the mean random error and the frequency (%) of replications in which error $\geq$ 0.1 (Table 1). For the position of the DIF-items, *pos*, a trend was observed for a higher bias when the DIF-items location parameters were uniformly distributed on the latent trait continuum (“Unif”) or when the DIF-items were the easiest ones of the scale (“Low”).

Table 1

Simulations without DIF

			Error	Error ≤ –0.1	Error ≥ 0.1
γ	*N*	*J*	Mean (*SD*)	*N* (%)	*N* (%)
0	100	4 items	0.001 (0.167)	139 (27.8)	138 (27.6)
		8 items	–0.001 (0.200)	161 (32.2)	156 (31.2)
	200	4 items	–0.001 (0.119)	98 (19.6)	94 (18.8)
		8 items	0.012 (0.136)	98 (19.6)	117 (23.4)
0.1	100	4 items	0.008 (0.168)	121 (24.2)	140 (28.0)
		8 items	–0.010 (0.203)	170 (34.0)	144 (28.8)
	200	4 items	–0.003 (0.114)	94 (18.8)	93 (18.6)
		8 items	–0.002 (0.145)	124 (24.8)	117 (23.4)
0.2	100	4 items	–0.007 (0.170)	143 (28.6)	132 (26.4)
		8 items	–0.002 (0.200)	151 (30.2)	144 (28.8)
	200	4 items	0.004 (0.121)	93 (18.6)	111 (22.2)
		8 items	0.006 (0.130)	102 (20.4)	120 (24.0)

Note. Mean random error, standard deviation (*SD*) and frequency (%) of replications (over 500) in which error ≥ 0.1 and in which error ≤ –0.1 in the case where no differential item functioning was simulated (*N*: group size, γ: difference in the mean latent trait level between groups, *J*: number of items in the scale)

When all entered into a multivariate model of analysis of variance, the six factors were significantly associated with bias (p-value < 0.0001). However, the estimated coefficients associated with group size N, number of items in the scale J, and mean latent trait level difference between groups γ were so small (i.e., absolute value <0.01) that they were considered to be negligible. For DIF-size δ_{dif} and the proportion of DIF-items p, the estimated coefficients were all higher than 0.1; consequently, we tested the significance of the interaction between these two factors (Table 5). Taking into account the significant effect of this interaction term, the mean increase of the mean bias was meaningful (>0.1) when the DIF-size δ_{dif} and the proportion of DIF-items p were at least: 0.25 and 75%, or 0.5 and 50%, or 1 and 25%, respectively. Concerning the position of the DIF-items, *pos*, the mean bias was higher when the DIF-items were the easiest ones or if the DIF-items location parameters were uniformly distributed on the latent trait continuum; however, the estimated coefficients associated with each of the categories of *pos* were smaller than 0.04.

Table 2A

Simulations with uniform DIF (DIF-size = 0.25, group size = 100)

			$\gamma = 0$		$\gamma = 0.1$		$\gamma = 0.2$	
J	*p*	*pos*	Bias Mean (*SD*)	Bias ≥ 0.1 *N* (%)	Bias Mean (*SD*)	Bias ≥ 0.1 *N* (%)	Bias Mean (*SD*)	Bias ≥ 0.1 *N* (%)
4 items	25%	Unif	0.07 (0.16)	224 (44.8)	0.07 (0.18)	218 (43.6)	0.05 (0.17)	192 (38.4)
		Mean	0.07 (0.17)	213 (42.6)	0.06 (0.17)	207 (41.4)	0.06 (0.17)	196 (39.2)
		Extreme	0.06 (0.17)	205 (41.0)	0.05 (0.18)	198 (39.6)	0.07 (0.17)	205 (41.0)
		High	0.06 (0.16)	208 (41.6)	0.06 (0.17)	209 (41.8)	0.05 (0.18)	186 (37.2)
		Low	0.06 (0.18)	194 (38.8)	0.07 (0.17)	215 (43.0)	0.07 (0.18)	223 (44.6)
	50%	Unif	0.15 (0.17)	307 (61.4)	0.13 (0.17)	288 (57.6)	0.12 (0.17)	284 (56.8)
		Mean	0.13 (0.17)	279 (55.8)	0.13 (0.16)	273 (54.6)	0.13 (0.17)	283 (56.6)
		Extreme	0.10 (0.17)	256 (51.2)	0.13 (0.16)	278 (55.6)	0.11 (0.17)	250 (50.0)
		High	0.12 (0.17)	279 (55.8)	0.12 (0.17)	269 (53.8)	0.12 (0.16)	274 (54.8)
		Low	0.14 (0.17)	296 (59.2)	0.13 (0.17)	286 (57.2)	0.13 (0.16)	287 (57.4)
	75%	Unif	0.19 (0.17)	349 (69.8)	0.20 (0.17)	360 (72.0)	0.19 (0.18)	340 (68.0)
		Mean	0.19 (0.18)	346 (69.4)	0.18 (0.16)	365 (73.0)	0.19 (0.16)	350 (70.0)
		Extreme	0.17 (0.17)	334 (66.8)	0.19 (0.17)	348 (69.6)	0.20 (0.18)	356 (71.2)
		High	0.20 (0.17)	363 (72.6)	0.21 (0.17)	370 (74.0)	0.19 (0.17)	351 (70.2)
		Low	0.20 (0.18)	350 (70.0)	0.18 (0.16)	339 (67.8)	0.19 (0.17)	349 (69.8)
8 items	25%	Unif	0.06 (0.21)	214 (42.8)	0.07 (0.20)	221 (44.2)	0.07 (0.20)	229 (45.8)
		Mean	0.07 (0.22)	210 (42.0)	0.06 (0.21)	202 (40.4)	0.07 (0.19)	227 (45.4)
		Extreme	0.04 (0.20)	191 (38.2)	0.04 (0.20)	193 (38.6)	0.06 (0.21)	208 (41.6)
		High	0.05 (0.19)	205 (41.0)	0.05 (0.20)	200 (40.0)	0.06 (0.21)	211 (42.2)
		Low	0.04 (0.19)	194 (38.8)	0.05 (0.20)	188 (37.6)	0.07 (0.21)	221 (44.2)
	50%	Unif	0.14 (0.20)	288 (57.6)	0.14 (0.20)	282 (56.4)	0.14 (0.21)	290 (58.0)
		Mean	0.14 (0.20)	291 (58.2)	0.12 (0.21)	260 (52.0)	0.13 (0.19)	282 (56.4)
		Extreme	0.09 (0.21)	253 (50.6)	0.12 (0.20)	271 (54.2)	0.10 (0.20)	240 (48.0)
		High	0.13 (0.20)	283 (56.6)	0.12 (0.20)	265 (53.0)	0.13 (0.21)	276 (55.2)
		Low	0.13 (0.21)	271 (54.2)	0.14 (0.21)	291 (58.2)	0.11 (0.20)	268 (53.6)
	75%	Unif	0.20 (0.22)	341 (68.2)	0.19 (0.20)	340 (68.0)	0.20 (0.21)	340 (68.0)
		Mean	0.18 (0.19)	339 (67.8)	0.19 (0.20)	332 (66.4)	0.18 (0.20)	326 (65.2)
		Extreme	0.19 (0.21)	335 (67.0)	0.18 (0.21)	324 (64.8)	0.18 (0.20)	316 (63.2)
		High	0.19 (0.21)	336 (67.2)	0.18 (0.19)	340 (68.0)	0.18 (0.20)	335 (67.0)
		Low	0.18 (0.20)	333 (66.6)	0.18 (0.21)	329 (65.8)	0.19 (0.19)	341 (68.2)

Note. Mean bias, standard deviation (*SD*) and frequency (%) of replications (over 500) in which bias ≥ 0.1 when DIF-size = 0.25, the group size = 100 and the differential item functioning (DIF) was uniform (γ: difference in the mean latent trait level between groups, *J*: number of items in the scale, *p*: % of DIF-items, *pos*: position of the DIF-items location parameters along the latent trait, Unif: uniformly distributed, Mean: around the mean of the items difficulties, Extreme: high and low difficulties, High: high difficulties, Low: low difficulties)

Simulations with non-uniform DIF

Description of the mean (*SD*) bias and frequency (%) of replications, among 500, in which bias ≥ 0.1 or in which bias ≥ –0.1 in the two cases of non-uniform DIF studied, gentle and steep slope, are shown in Table 6 and 7 respectively. No clear trend concerning the three factors manipulated in each case of non-uniform DIF emerged: the mean bias was rarely higher than 0.01 or below –0.01 and the proportions of replications in which bias ≥ 0.1 or in which bias ≥ –0.1 were rarely higher than 30%. These values are very similar to those for random error (Table 1). The three DIF-related factors were entered together into a multivariate model of analysis of variance (one for each kind of non-uniform DIF studied): no statistically significant effect was found concerning the DIF-size δ_{dif} or the proportion of DIF-items *p*; however, there was a significant relationship for the position of the DIF-items, *pos*, in both kinds of non-uniform DIF studied (Table 8). The highest estimated coefficients (absolute values > 0.01) were associated with the "High" and "Low" categories and were of opposite sign for these two categories and also for the two kinds of non-uniform DIF studied.

Table 3A

Simulations with uniform DIF (DIF-size = 0.25, group size = 200)

			$\gamma = 0$		$\gamma = 0.1$		$\gamma = 0.2$	
J	*p*	*pos*	Bias Mean (*SD*)	Bias ≥ 0.1 *N* (%)	Bias Mean (*SD*)	Bias ≥ 0.1 *N* (%)	Bias Mean (*SD*)	Bias ≥ 0.1 *N* (%)
4 items	25%	Unif	0.07 (0.12)	204 (40.8)	0.07 (0.12)	204 (40.8)	0.06 (0.12)	177 (35.4)
		Mean	0.07 (0.12)	186 (37.2)	0.07 (0.11)	192 (38.4)	0.06 (0.11)	169 (33.8)
		Extreme	0.05 (0.11)	165 (33.0)	0.05 (0.12)	158 (31.6)	0.05 (0.12)	168 (33.6)
		High	0.05 (0.12)	166 (33.2)	0.05 (0.12)	168 (33.6)	0.06 (0.11)	173 (34.6)
		Low	0.06 (0.12)	187 (37.4)	0.07 (0.12)	200 (40.0)	0.06 (0.12)	194 (38.8)
	50%	Unif	0.14 (0.12)	311 (62.2)	0.13 (0.12)	304 (60.8)	0.14 (0.12)	307 (61.4)
		Mean	0.13 (0.12)	291 (58.2)	0.11 (0.12)	276 (55.2)	0.12 (0.11)	275 (55.0)
		Extreme	0.11 (0.12)	285 (57.0)	0.11 (0.12)	287 (57.4)	0.13 (0.12)	298 (59.6)
		High	0.12 (0.11)	279 (55.8)	0.12 (0.11)	281 (56.2)	0.12 (0.12)	285 (57.0)
		Low	0.12 (0.12)	275 (55.0)	0.12 (0.12)	281 (56.2)	0.12 (0.12)	278 (55.6)
	75%	Unif	0.19 (0.12)	396 (79.2)	0.20 (0.12)	393 (78.6)	0.19 (0.12)	386 (77.2)
		Mean	0.17 (0.12)	370 (74.0)	0.19 (0.12)	389 (77.8)	0.19 (0.11)	382 (76.4)
		Extreme	0.18 (0.12)	375 (75.0)	0.18 (0.12)	383 (76.6)	0.18 (0.12)	370 (74.0)
		High	0.19 (0.11)	387 (77.4)	0.19 (0.12)	388 (77.6)	0.20 (0.13)	391 (78.2)
		Low	0.19 (0.12)	372 (74.4)	0.19 (0.12)	389 (77.8)	0.19 (0.12)	388 (77.6)
8 items	25%	Unif	0.08 (0.14)	228 (45.6)	0.06 (0.14)	193 (38.6)	0.05 (0.15)	180 (36.0)
		Mean	0.07 (0.14)	216 (43.2)	0.06 (0.14)	184 (36.8)	0.06 (0.14)	194 (38.8)
		Extreme	0.06 (0.14)	204 (40.8)	0.05 (0.15)	188 (37.6)	0.05 (0.14)	171 (34.2)
		High	0.05 (0.14)	176 (35.2)	0.05 (0.14)	182 (37.6)	0.04 (0.15)	167 (33.4)
		Low	0.05 (0.14)	185 (37.0)	0.05 (0.14)	184 (36.8)	0.06 (0.15)	198 (39.6)
	50%	Unif	0.13 (0.13)	302 (60.4)	0.13 (0.15)	288 (57.6)	0.14 (0.14)	310 (62.0)
		Mean	0.12 (0.15)	272 (54.4)	0.13 (0.14)	278 (55.6)	0.12 (0.15)	290 (58.0)
		Extreme	0.12 (0.14)	270 (54.0)	0.10 (0.13)	245 (49.0)	0.10 (0.13)	248 (49.6)
		High	0.12 (0.14)	278 (55.6)	0.12 (0.13)	280 (56.0)	0.10 (0.14)	253 (50.6)
		Low	0.14 (0.14)	308 (61.6)	0.13 (0.15)	287 (57.4)	0.13 (0.14)	295 (59.0)
	75%	Unif	0.19 (0.15)	364 (72.8)	0.20 (0.14)	373 (74.6)	0.19 (0.13)	366 (73.2)
		Mean	0.17 (0.14)	360 (72.0)	0.18 (0.14)	357 (71.4)	0.18 (0.14)	356 (71.2)
		Extreme	0.17 (0.13)	349 (69.8)	0.18 (0.14)	361 (72.2)	0.17 (0.14)	352 (70.4)
		High	0.20 (0.14)	380 (76.0)	0.20 (0.14)	378 (75.6)	0.20 (0.14)	379 (75.8)
		Low	0.20 (0.14)	392 (78.4)	0.19 (0.14)	379 (75.8)	0.19 (0.14)	377 (75.4)

Note. Mean bias, standard deviation (*SD*) and frequency (%) of replications (over 500) in which bias ≥ 0.1 when DIF-size = 0.25, the group size = 200 and the differential item functioning (DIF) was uniform (γ: difference in the mean latent trait level between groups, *J*: number of items in the scale, *p*: % of DIF-items, *pos*: position of the DIF-items location parameters along the latent trait, Unif: uniformly distributed, Mean: around the mean of the items difficulties, Extreme: high and low difficulties, High: high difficulties, Low: low difficulties)

Discussion

This simulation study addressed the issue of the practical implications of the presence of DIF-item(s) within a scale used in realistic conditions of health research studies, to measure a difference between two groups concerning a construct evaluated on the latent trait scale using a PCM. If the presence of uniform DIF was not taken into account, the mean bias affecting the difference between the two groups was higher than 0.03 whatever the simulated conditions; this is more than double the highest absolute value of the mean random error found in the simulations without DIF. However, in the multivariate analysis, three of the six factors which were manipulated in the simulation model were found to have negligible effects (absolute value <0.01): the group size *N*, the number of items in the scale *J*, and the mean latent trait level difference between groups γ. Two factors were found to have meaningful effects (absolute value >0.1, as defined in this study) and also to be involved in an interaction: the DIF-size, δ_{dif}, and the proportion of DIF-items, *p*. In practical terms, this interaction means that the effect of the size of uniform DIF increased with the number of DIF-affected items in the scale. Therefore, the definition of a meaningful DIF cannot be based only on the DIF-size, as is the case if indices of DIF effect-size recommended in the literature (0.5 logit or ETS rules) are used. Indeed, this simulation study shows that a DIF-size of 0.25 (considered as negligible according to the 0.5 logit or ETS rules) affecting 75% of the items of the scale could have a meaningful effect as, in this configuration, the mean bias on the estimated difference between groups was higher than 0.1.

The last factor studied in the case of uniform DIF was the position of the DIF-items location parameters along the latent trait, *pos*. The "Unif" category, in which the DIF-items location parameters were uniformly distributed along the latent trait continuum, was chosen as the reference

Table 4

Simulations with uniform DIF (group size = 200, number of items in the scale = 8, mean latent trait level equal in the two groups)

		δ_{dif} = 0.25		δ_{dif} = 0.5		δ_{dif} = 1	
		Bias	Bias ≥ 0.1	Bias	Bias ≥ 0.1	Bias	Bias ≥ 0.1
p	*pos*	Mean (*SD*)	*N* (%)	Mean (*SD*)	*N* (%)	Mean (*SD*)	*N* (%)
25%	Unif	0.08 (0.14)	228 (45.6)	0.14 (0.14)	299 (59.8)	0.25 (0.13)	433 (86.6)
	Mean	0.07 (0.14)	216 (43.2)	0.13 (0.14)	291 (58.2)	0.24 (0.13)	432 (86.4)
	Extreme	0.06 (0.14)	204 (40.8)	0.10 (0.14)	250 (50.0)	0.18 (0.13)	369 (73.8)
	High	0.05 (0.14)	176 (35.2)	0.09 (0.13)	233 (46.6)	0.16 (0.13)	345 (69.0)
	Low	0.05 (0.14)	185 (37.0)	0.11 (0.14)	264 (52.8)	0.23 (0.14)	403 (80.6)
50%	Unif	0.13 (0.13)	302 (60.4)	0.26 (0.13)	447 (89.4)	0.49 (0.12)	500 (100.0)
	Mean	0.12 (0.15)	272 (54.4)	0.24 (0.14)	422 (84.4)	0.43 (0.13)	496 (99.2)
	Extreme	0.12 (0.14)	270 (54.0)	0.22 (0.13)	416 (83.2)	0.39 (0.13)	492 (98.4)
	High	0.12 (0.14)	278 (55.6)	0.23 (0.14)	419 (83.8)	0.42 (0.12)	497 (99.4)
	Low	0.14 (0.14)	308 (61.6)	0.26 (0.13)	443 (88.6)	0.48 (0.13)	500 (100.0)
75%	Unif	0.19 (0.15)	364 (72.8)	0.39 (0.14)	490 (98.0)	0.75 (0.13)	500 (100.0)
	Mean	0.17 (0.14)	360 (72.0)	0.36 (0.14)	483 (96.6)	0.69 (0.13)	500 (100.0)
	Extreme	0.17 (0.13)	349 (69.8)	0.34 (0.14)	479 (95.8)	0.65 (0.13)	500 (100.0)
	High	0.20 (0.14)	380 (76.0)	0.37 (0.14)	491 (98.2)	0.70 (0.13)	500 (100.0)
	Low	0.20 (0.14)	392 (78.4)	0.39 (0.14)	489 (97.8)	0.74 (0.14)	500 (100.0)

Note. Mean bias, standard deviation (*SD*) and frequency (%) of replications (over 500) in which bias ≥ 0.1 in the case of uniform differential item functioning (DIF) when the group size was set at 200, the number of items in the scale was set at 8 and there was no difference in the mean latent trait level between the groups (*p*: proportion of items with DIF, *pos*: position of the DIF-items location parameters along the latent trait, Unif: uniformly distributed, Mean: around the mean of the item difficulties, Extreme: high and low difficulties, High: high difficulties, Low: low difficulties) category in the multivariate model of analysis of variance. In this model, the absolute values of the estimated coefficients were higher than 0.01, but lower than 0.1, for the "Mean" category (DIF-items location parameters close to the mean of the item difficulties in the scale), the "Extreme" category (DIF-items were the easiest and most difficult ones in the scale) and the "High" category (DIF-items were the most difficult ones). This means for example that, when the DIF-items were the easiest ones or were uniformly distributed, there was a slightly but not a meaningfully higher mean bias than in the case in which the DIF-items were the highest ones. This may indicate that, in addition to the DIF-size, δ_{dif}, and the proportion of DIF-items p, their level of difficulty should be taken into account when considering the practical meaning of DIF-items within a scale. Moreover, the level of difficulty of the DIF-items was the only one which was found to be significantly associated with the bias in both multivariate models of analysis of variance (gentle and steep slope), in the case of non-uniform DIF. With reference to the "Unif" category (DIF-items location parameters uniformly distributed along the latent trait continuum), if the DIF-items were the

Table 5

Results of the multivariate analysis of variance in the case of uniform DIF

Factor	Category	Estimate (Standard Error)	Degrees of freedom	Test statistic	*p*-value
N	100	Reference	1	23.52	<0.0001
	200	–0.003 (0.001)			
J	4	Reference	1	21.17	<0.0001
	8	–0.003 (0.001)			
γ	0	Reference	2	32.02	<0.0001
	0.1	–0.003 (0.001)			
	0.2	–0.006 (0.001)			
δ_{dif}	0.25	Reference	2	1.10^5	<0.0001
	0.5	0.056 (0.001)			
	1	0.152 (0.001)			
p	25%	Reference	2	77197.96	<0.0001
	50%	0.065 (0.001)			
	75%	0.129 (0.001)			
pos	Unif	Reference	4	10184.61	<0.0001
	Mean	–0.018 (0.001)			
	Extreme	–0.039 (0.001)			
	High	–0.029 (0.001)			
	Low	–0.004 (0.001)			
Interaction	0.5*50%	0.061 (0.002)	4	592.01	<0.0001
δ_{dif} * *p*	0.5*75%	0.125 (0.002)			
	1*50%	0.169 (0.002)			
	1*75%	0.360 (0.002)			
Constant		0.053 (0.001)	1	66.29	<0.0001

Note. *N*: group size, γ: difference in the mean latent trait level between groups, *J*: number of items in the scale, *p*: proportion of items with DIF, *pos*: position of the DIF-items location parameters along the latent trait, Unif: uniformly distributed, Mean: around the mean of the item difficulties, Extreme: high and low difficulties, High: high difficulties, Low: low difficulties

Table 6

Simulations with gentle-slope non-uniform DIF (group size = 200, number of items in the scale = 8, difference in the mean latent trait level between groups = 0.1)

		$\delta_{dif} = 0.25$			$\delta_{dif} = 0.5$			$\delta_{dif} = 1$		
p	*pos*	Bias Mean (*SD*)	Bias ≤ –0.1 *N* (%)	Bias ≥ 0.1 *N* (%)	Bias Mean (*SD*)	Bias ≤ –0.1 *N* (%)	Bias ≥ 0.1 *N* (%)	Bias Mean (*SD*)	Bias ≤ –0.1 *N* (%)	Bias ≥ 0.1 *N* (%)
25%	Unif	–0.016 (0.14)	133 (26.6)	103 (20.6)	–0.001 (0.14)	122 (24.4)	125 (25.0)	–0.001 (0.14)	116 (23.2)	117 (23.4)
	Mean	0.002 (0.14)	117 (23.4)	131 (26.2)	–0.001 (0.14)	122 (24.4)	117 (23.4)	–0.004 (0.14)	122 (24.4)	113 (22.6)
	Extreme	–0.001 (0.13)	115 (23.0)	119 (23.8)	–0.006 (0.14)	120 (24.0)	114 (22.8)	0.006 (0.14)	114 (22.8)	127 (25.4)
	High	–0.008 (0.13)	123 (24.6)	103 (20.6)	–0.018 (0.14)	144 (28.8)	103 (20.6)	–0.026 (0.13)	144 (28.8)	82 (16.4)
	Low	0.015 (0.14)	109 (21.8)	136 (27.2)	0.014 (0.14)	101 (20.2)	124 (24.8)	0.033 (0.15)	85 (17.0)	162 (32.4)
50%	Unif	–0.012 (0.14)	127 (25.4)	119 (23.8)	–0.007 (0.13)	108 (21.6)	100 (20.0)	–0.004 (0.13)	127 (25.4)	117 (23.4)
	Mean	–0.012 (0.13)	122 (24.4)	107 (21.4)	–0.007 (0.13)	121 (24.2)	107 (21.4)	–0.001 (0.13)	112 (22.4)	107 (21.4)
	Extreme	–0.008 (0.14)	113 (22.6)	97 (19.4)	–0.015 (0.14)	139 (27.8)	96 (19.2)	–0.005 (0.15)	128 (25.6)	115 (23.0)
	High	0.001 (0.13)	116 (23.2)	114 (22.8)	–0.017 (0.14)	126 (25.2)	103 (20.6)	–0.031 (0.14)	153 (30.6)	88 (17.6)
	Low	–0.001 (0.14)	125 (25.0)	116 (23.2)	0.013 (0.14)	106 (21.2)	136 (27.2)	0.033 (0.14)	86 (17.2)	157 (31.4)
75%	Unif	0.002 (0.14)	113 (22.6)	119 (23.8)	0.006 (0.14)	115 (23.0)	121 (24.2)	0.002 (0.14)	117 (23.4)	108 (21.6)
	Mean	–0.004 (0.14)	118 (23.6)	112 (22.4)	0.005 (0.14)	115 (23.0)	112 (22.4)	–0.001 (0.14)	115 (23.0)	115 (23.0)
	Extreme	0.004 (0.13)	117 (23.4)	125 (25.0)	0.001 (0.13)	104 (20.8)	113 (22.6)	–0.007 (0.13)	130 (26.0)	102 (20.4)
	High	–0.004 (0.14)	110 (22.0)	103 (20.6)	–0.013 (0.14)	128 (25.6)	108 (21.6)	–0.027 (0.13)	149 (29.8)	80 (16.0)
	Low	0.003 (0.14)	108 (21.6)	115 (23.0)	0.006 (0.14)	117 (23.4)	118 (23.6)	0.022 (0.13)	91 (18.2)	140 (28.0)

Note. Mean bias, standard deviation (*SD*) and frequency (%) of replications (over 500) in which bias ≥ 0.1 and in which bias ≤ –0.1 in the cases of the gentle slope non-uniform differential item functioning (DIF) when the group size was set at 200, the number of items in the scale at 8 and the difference in the mean latent trait level between groups was set at 0.1 (*p*: proportion of DIF-items, *pos*: position of the DIF-items location parameters along the latent trait, Unif: uniformly distributed, Mean: around the mean of the item difficulties, Extreme: high and low difficulties, High: high difficulties, Low: low difficulties)

Table 7.

Simulations with steep-slope non-uniform DIF (group size = 200, number of items in the scale = 8, difference in the mean latent trait level between groups = 0.1)

		δ_{dif} = 0.25			δ_{dif} = 0.5			δ_{dif} = 1		
p	*pos*	Bias Mean (*SD*)	Bias ≤ –0.1 *N* (%)	Bias ≥ 0.1 *N* (%)	Bias Mean (*SD*)	Bias ≤ –0.1 *N* (%)	Bias ≥ 0.1 *N* (%)	Bias Mean (*SD*)	Bias ≤ –0.1 *N* (%)	Bias ≥ 0.1 *N* (%)
25%	Unif	0.003 (0.15)	119 (23.8)	131 (26.2)	–0.003 (0.14)	111 (22.2)	120 (24.0)	–0.005 (0.14)	132 (26.4)	120 (24.0)
	Mean	–0.001 (0.15)	133 (26.6)	123 (24.6)	–0.003 (0.13)	113 (22.6)	107 (21.4)	0.002 (0.15)	128 (25.6)	128 (25.6)
	Extreme	–0.001 (0.15)	121 (24.2)	128 (25.6)	–0.008 (0.14)	124 (24.8)	119 (23.8)	0.007 (0.13)	110 (22.0)	125 (25.0)
	High	–0.001 (0.15)	128 (25.6)	137 (27.4)	–0.005 (0.14)	126 (25.2)	116 (23.2)	0.023 (0.14)	97 (19.4)	144 (28.8)
	Low	–0.005 (0.14)	130 (26.0)	112 (22.4)	–0.027 (0.13)	135 (27.0)	85 (17.0)	–0.030 (0.14)	145 (29.0)	89 (17.8)
50%	Unif	–0.013 (0.14)	130 (26.0)	105 (21.0)	0.013 (0.14)	104 (20.8)	132 (26.4)	–0.004 (0.15)	139 (27.8)	119 (23.8)
	Mean	0.005 (0.14)	112 (22.4)	131 (26.2)	–0.004 (0.14)	121 (24.2)	110 (22.0)	0.004 (0.15)	123 (24.6)	118 (23.6)
	Extreme	–0.002 (0.14)	112 (22.4)	111 (22.2)	–0.009 (0.14)	134 (26.8)	109 (21.8)	0.001 (0.14)	116 (23.2)	119 (23.8)
	High	0.007 (0.14)	101 (20.2)	125 (25.0)	0.023 (0.13)	97 (19.4)	151 (30.2)	0.026 (0.14)	99 (19.8)	156 (31.2)
	Low	–0.008 (0.15)	131 (26.2)	121 (24.2)	–0.003 (0.15)	126 (25.2)	125 (25.0)	–0.038 (0.15)	172 (34.4)	94 (18.8)
75%	Unif	–0.004 (0.15)	128 (25.6)	121 (24.2)	–0.008 (0.13)	124 (24.8)	112 (22.4)	0.001 (0.15)	124 (24.8)	132 (26.4)
	Mean	0.006 (0.15)	116 (23.2)	130 (26.0)	–0.001 (0.15)	122 (24.4)	123 (24.6)	–0.019 (0.15)	152 (30.4)	106 (21.2)
	Extreme	0.003 (0.14)	117 (23.4)	118 (23.6)	–0.011 (0.14)	140 (28.0)	110 (22.0)	0.005 (0.13)	97 (19.4)	108 (21.6)
	High	0.008 (0.14)	106 (21.2)	127 (25.4)	0.019 (0.14)	94 (18.8)	128 (25.6)	0.031 (0.15)	96 (19.2)	168 (33.6)
	Low	–0.013 (0.14)	135 (27.0)	99 (19.8)	–0.007 (0.14)	131 (26.2)	112 (22.4)	–0.023 (0.14)	141 (28.2)	93 (18.6)

Note. Mean bias, standard deviation (*SD*) and frequency (%) of replications (over 500) in which bias ≥ 0.1 and in which bias ≤ –0.1 in the cases of the steep slope non-uniform differential item functioning (DIF) when the group size was set at 200, the number of items in the scale at 8 and the difference in the mean latent trait level between groups was set at 0.1 (*p*: proportion of DIF-items, *pos*: position of the DIF-items location parameters along the latent trait, Unif: uniformly distributed, Mean: around the mean of the item difficulties, Extreme: high and low difficulties, High: high difficulties, Low: low difficulties)

Table 8

Results of the multivariate analyses of variance in the two cases of non-uniform DIF studied (gentle and steep slope)

		Gentle slope				Steep slope			
Factor	Category	Estimate (Standard Error)	Degrees of freedom	Test statistic	*p*-value	Estimate (Standard Error)	Degrees of freedom	Test statistic	*p*-value
δ_{dif}	0.25	Reference	2	0.36	0.696	Reference	2	0.21	0.809
	0.5	0.0001 (0.0023)				–0.0014 (0.0023)			
	1	0.0016 (0.0023)				–0.0003 (0.0023)			
p	25%	Reference	2	2.55	0.078	Reference	2	1.01	0.364
	50%	–0.0042 (0.0023)				0.0032 (0.0023)			
	75%	–0.0004 (0.0023)				0.0023 (0.0023)			
pos	Unif	Reference	4	29.22	<0.0001	Reference	4	28.3	<0.0001
	Mean	0.0008 (0.0029)				0.0012 (0.0030)			
	Extreme	–0.0002 (0.0029)				0.0005 (0.0030)			
	High	–0.0125 (0.0029)				0.0170 (0.0030)			
	Low	0.0186 (0.0029)				–0.0149 (0.0030)			
Constant		–0.0026 (0.0028)	1	0.95	0.343	–0.0035 (0.0028)	1	1.21	0.225

Note. δ_{dif}: DIF-size, *p*: proportion of DIF-items, *pos*: position of the DIF-items location parameters along the latent trait, Unif: uniformly distributed, Mean: around the mean of the item difficulties, Extreme: high and low difficulties, High: high difficulties, Low: low difficulties easiest ones, the mean bias was slightly greater in the gentle slope non-uniform DIF and slightly smaller in the case of the steep slope non-uniform DIF. These effects were reversed if the DIF-items were the most difficult ones. The influence of the position of the DIF-items, *pos*, on the mean bias, although statistically significant, was nevertheless marginal.

Indeed, the mean bias observed in the combinations studied in the case of non-uniform DIF was not very different from the mean random error observed in the simulations without DIF. This may be because the DIF-sizes used in these simulations were not sufficiently large to cause meaningful bias at the scale level or because the non-uniform DIF simulated in this study did not modify the item difficulty. Indeed, in the simulation model, the sum of the sizes of the DIF affecting each response category location parameter was equal to zero for each DIF-item; this may have resulted in a phenomenon at the item level similar to what is currently called "DIF cancellation" at the scale level, i.e., when an item or a set of items exhibiting DIF for one group cancels the effects associated with other items that exhibit DIF against another group such that there is no differential functioning at the test level (Nandakumar, 1993; Shealy and Stout, 1993; Teresi, 2006; Wyse, 2013). Further studies are needed to explore the effect of non-uniform DIF, especially that responsible for modification of item difficulty. It would be valuable to investigate extended values of the magnitude of the mean latent trait level difference between the two groups γ as the values studied here represented small differences between the two groups. Indeed, the minimal clinically important difference of a questionnaire is sometimes defined as being equal to half a standard deviation of the score, and this is higher than the γ values studied in this work (Norman, Sloan, and Wyrwich, 2003). Also, it would be interesting to investigate the consequences of the two groups compared being of different sizes, as is frequent in practice in health research studies. Finally, the PCM was used in this simulation study, so other studies are required to evaluate the effect of DIF when other models are used, for example, the factor common model or item response theory models.

Some important and new information nevertheless emerges from this simulation study and can be used to formulate recommendations concerning the presence of DIF-items in health-related questionnaires. Indeed, the 0.5 logit or ETS rules are not sufficient to evaluate the meaning (or implications) of the presence of DIF at the scale level. In addition, the percentage of items of the scale which are affected by DIF needs to be taken into account, as has, to a lesser extent, the level of difficulty of the DIF-items. This study shows that, in practice, if less than 50% of the items of the scale are affected by a uniform DIF whose size is smaller than 0.25, whatever the level of difficulty of these DIF-items, the resulting measurement bias at the scale level is not likely to be meaningful. Moreover, if the DIF is non-uniform and does not modify the item difficulty, its effect at the scale level may be considered to be negligible in the studied conditions of questionnaire use in health research.

References

Banh, M. K., Crane, P. K., Rhew, I., Gudmundsen, G., Stoep, A. V., Lyon, A., et al. (2012). Measurement equivalence across racial/ethnic groups of the mood and feelings questionnaire for childhood depression. *Journal of Abnormal Child Psychology*, *40*, 353-367.

Barbier, M., Peters, S., and Hansez, I. (2009). Measuring positive and negative occupational states (PNOSI): Structural confirmation of a new Belgian tool. *Psychologica Belgica*, *49*.

Bjorner, J. B., Kreiner, S., Ware, J. E., Damsgaard, M. T., and Bech, P. (1998). Differential item functioning in the Danish translation of the SF-36. *Journal of Clinical Epidemiology*, *51*, 1189-1202.

Bjorner, J. B., and Pejtersen, J. H. (2010). Evaluating construct validity of the second version of the Copenhagen Psychosocial Questionnaire through analysis of differential item functioning and differential item effect. *Scandinavian Journal of Public Health*, *38*, 90-105.

Borsboom, D. (2006). When does measurement invariance matter? [Editorial]. *Medical Care Measurement in a Multi-Ethnic Society*, *44*, 176-181.

Christensen, K. B., Kreiner, S., and Mesbah, M. (Eds.). (2012). Front matter. In *Rasch models in health* (pp. i-xvi). Hoboken, NJ: John Wiley.

Coste, J., Tissier, F., Pouchot, J., Ecosse, E., Rouquette, A., Bertagna, X., et al. (2014). Rasch analysis for assessing unidimensionality and identifying measurement biases of malignancy scores in oncology. The example of the Weiss histopathological system for the diagnosis of adrenocortical cancer. *Cancer Epidemiology*, *38*, 200-208.

Crane, P. K., Gibbons, L. E., Ocepek-Welikson, K., Cook, K., Cella, D., Narasimhalu, K., et al. (2007). A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *16 Suppl 1*, 69-84.

Crane, P. K., Gibbons, L. E., Willig, J. H., Mugavero, M. J., Lawrence, S. T., Schumacher, J. E., et al. (2010). Measuring depression levels in HIV-infected patients as part of routine clinical care using the nine-item Patient Health Questionnaire (PHQ-9). *AIDS Care*, *22*, 874-885.

Crane, P. K., Hart, D. L., Gibbons, L. E., and Cook, K. F. (2006). A 37-item shoulder functional status item pool had negligible differential item functioning. *Journal of Clinical Epidemiology*, *59*, 478-484.

Crane, P. K., Narasimhalu, K., Gibbons, L. E., Pedraza, O., Mehta, K. M., Tang, Y., et al. (2008). Composite scores for executive function items: Demographic heterogeneity and relationships with quantitative magnetic resonance imaging. *Journal of the International Neuropsychological Society: JINS*, *14*, 746-759.

Czachowski, S., Terluin, B., Izdebski, A., and Izdebski, P. (2012). Evaluating the cross-cultural validity of the Polish version of the Four-Dimensional Symptom Questionnaire (4DSQ) using differential item functioning (DIF) analysis. *Family Practice*, *29*, 609-615.

Dorans, N. J., and Holland, P. W. (1992). DIF detection and description: Mantel-Haenszel and Standardization 1,2. *ETS Research Report Series*, *1992*, 1-40.

Earleywine, M., LaBrie, J. W., and Pedersen, E. R. (2008). A brief Rutgers Alcohol Problem Index with less potential for bias. *Addictive Behaviors*, *33*, 1249-1253.

Fleishman, J. A., Spector, W. D., and Altman, B. M. (2002). Impact of differential item functioning on age and gender differences in functional disability. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*, *57*, S275-S284.

Gibbons, L. E., McCurry, S., Rhoads, K., Masaki, K., White, L., Borenstein, A. R., et al. (2009). Japanese-English language equivalence of the Cognitive Abilities Screening Instrument among Japanese-Americans. *International Psychogeriatrics / IPA*, *21*, 129-137.

Goetz, C., Ecosse, E., Rat, A.-C., Pouchot, J., Coste, J., and Guillemin, F. (2011). Measurement properties of the osteoarthritis of knee and hip quality of life OAKHQOL questionnaire: An item response theory analysis. *Rheumatology*, *50*, 500-505.

Golia, S. (2010). The assessment of DIF on Rasch measures with an application to job satisfaction. *Electronic Journal of Applied Statistical Analysis: Decision Support Systems and Services Evaluation*, *1*, 16-25.

Hardouin, J. B., Audureau, E., Leplège, A., and Coste, J. (2012). Spatio-temporal Rasch analysis of quality of life outcomes in the French general population. Measurement invariance and group comparisons. *BMC Medical Research Methodology*, *12*, 182.

Hart, D. L., Deutscher, D., Crane, P. K., and Wang, Y.-C. (2009). Differential item functioning was negligible in an adaptive test of functional status for patients with knee impairments who spoke English or Hebrew. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *18*, 1067-1083.

Jarl, G. M., Heinemann, A. W., and Norling Hermansson, L. M. (2012). Validity evidence for a modified version of the Orthotics and Prosthetics Users' Survey. *Disability and Rehabilitation: Assistive Technology*, *7*, 469-478.

Jones, R. N. (2003). Racial bias in the assessment of cognitive functioning of older adults. *Aging and Mental Health*, *7*, 83-102.

Jones, R. N., and Gallo, J. J. (2002). Education and sex differences in the Mini-Mental State Examination effects of differential item functioning. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *57*, 548-558.

Kenaszchuk, C., Wild, T. C., Rush, B. R., and Urbanoski, K. (2013). Rasch model of the GAIN substance problem scale among Canadian adults seeking residential and outpatient addiction treatment. *Addictive Behaviors*, *38*, 2279-2287.

King, M. W., Street, A. E., Gradus, J. L., Vogt, D. S., and Resick, P. A. (2013). Gender differences in posttraumatic stress symptoms among OEF/OIF veterans: An item response theory analysis. *Journal of Traumatic Stress*, *26*, 175-183.

Lai, J., Cella, D., Chang, C.-H., Bode, R. K., and Heinemann, A. W. (2003). Item banking to improve, shorten and computerize self-reported fatigue: An illustration of steps to create a core item bank from the FACIT-Fatigue Scale. *Quality of Life Research*, *12*, 485-501.

Lange, R., Thalbourne, M. A., Houran, J., and Lester, D. (2002). Depressive response sets due to gender and culture-based differential item functioning. *Personality and Individual Differences*, *33*, 937-954.

Lee, Y.-H., and Zhang, J. (2010). *Differential item functionning: Its consequences*. ETS Research Report (No. RR-10-01). Princeton, NJ: Educational Testing Service

Li, Z., and Zumbo, B., D. (2009). Impact of differential item functionning on subsequent statistical conclusion based on observed test score data. *Psicologica*, *30*, 343-370.

Linacre, J. M. (1994). Sample size and item calibration [or person measure] stability. *Rasch Measurement Transactions*, *7*, 328.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *The Journal of Applied Psychology*, *95*, 728-743.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127-143.

Meredith, W., and Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, *44*(11), S69-S77.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Millsap, R. E., and Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*, 297-334.

Morales, L. S., Reise, S. P., and Hays, R. D. (2000). Evaluating the equivalence of health care ratings by whites and Hispanics. *Medical Care*, *38*, 517-527.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's Test for DIF. *Journal of Educational Measurement*, *30*, 293-311.

Norman, G. R., Sloan, J. A., and Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Medical Care*, *41*, 582-592.

Orlando, M., and Marshall, G. N. (2002). Differential item functioning in a Spanish translation of the PTSD Checklist: Detection and evaluation of impact. *Psychological Assessment*, *14*, 50–59.

Paek, I., and Wilson, M. (2011). Formulating the Rasch differential item functioning model under the marginal maximum likelihood estimation context and its comparison with Mantel–Haenszel procedure in short test and small sample conditions. *Educational and Psychological Measurement*, *71*, 1023-1046.

Petersen, M. A., Groenvold, M., Aaronson, N. K., Chie, W.-C., Conroy, T., Costantini, A., et al. (2010). Development of computerised adaptive testing (CAT) for the EORTC QLQ-C30 dimensions – General approach and initial results for physical functioning. *European Journal of Cancer*, *46*, 1352-1358.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago, IL: University of Chicago Press.)

Rodriguez, H. P., and Crane, P. K. (2011). Examining multiple sources of differential item functioning on the Clinician and Group CAHPS® Survey. *Health Services Research*, *46*, 1778-1802.

Sébille, V., Blanchin, M., Guillemin, F., Falissard, B., and Hardouin, J.-B. (2014). A simple ratio-based approach for power and sample size determination for 2-group comparison using Rasch models. *BMC Medical Research Methodology*, *14*, 87.

Shealy, R., and Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*, 159-194.

Song, H., Cai, H., Brown, J. D., and Grimm, K. J. (2011). Differential item functioning of the Rosenberg Self-Esteem Scale in the US and China: Measurement bias matters. *Asian Journal of Social Psychology*, *14*, 176-188.

StataCorp, L. P. (2012). Stata statistical software: Release 12.1 [Computer software]. College Station, TX: StataCorp, L. P.

Steinberg, L., and Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods*, *11*, 402-415.

Tennant, A., and Pallant, J. F. (2007). DIF matters: A practical approach to test if differential item functioning makes a difference. *Rasch Measurement Transactions*, *20*, 1082-1084.

Teresi, J. A. (2006). Different approaches to differential item functioning in health applications. Advantages, disadvantages and some neglected topics. *Medical Care*, *44*, S152-170.

Teresi, J. A., Ramirez, M., Jones, R. N., Choi, S., and Crane, P. K. (2012). Modifying measures based on differential item functioning (DIF) impact analyses. *Journal of Aging and Health*, *24*, 1044-1076.

Tristan, A. (2006). An adjustment for sample size in DIF analysis. *Rasch Measurement Transactions*, *20*, 1070.

Wanders, R. B. K., Wardenaar, K. J., Kessler, R. C., Penninx, B. W. J. H., Meijer, R. R., and de Jonge, P. (2015). Differential reporting of depressive symptoms across distinct clinical subpopulations: What DIFference does it make? *Journal of Psychosomatic Research*, *78*, 130-136.

Woodbury, M. L., Velozo, C. A., Richards, L. G., Duncan, P. W., Studenski, S., and Lai, S.-M. (2008). Longitudinal stability of the Fugl-Meyer Assessment of the upper extremity. *Archives of Physical Medicine and Rehabilitation*, *89*, 1563-1569.

Wyse, A. E. (2013). DIF cancellation in the Rasch Model. *Journal of Applied Measurement*, *14*, 118-128.

Yu, Y. F., Yu, A. P., and Ahn, J. (2007). Investigating differential item functioning by chronic diseases in the SF-36 health survey: A latent trait analysis using MIMIC models. *Medical Care*, *45*, 851-859.

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

Appendix

Table 2B

Simulations with uniform DIF (DIF-size = 0.5, group size = 100)

			$\gamma = 0$		$\gamma = 0.1$		$\gamma = 0.2$	
J	*p*	*pos*	Bias Mean (*SD*)	Bias ≥ 0.1 *N* (%)	Bias Mean (*SD*)	Bias ≥ 0.1 *N* (%)	Bias Mean (*SD*)	Bias ≥ 0.1 *N* (%)
4 items	25%	Unif	0.13 (0.17)	275 (55.0)	0.12 (0.16)	278 (55.6)	0.13 (0.17)	280 (56.0)
		Mean	0.13 (0.17)	283 (56.6)	0.13 (0.17)	286 (57.2)	0.12 (0.17)	289 (57.8)
		Extreme	0.11 (0.17)	257 (51.4)	0.12 (0.17)	256 (51.2)	0.11 (0.16)	259 (51.8)
		High	0.11 (0.18)	248 (49.6)	0.10 (0.17)	245 (49.0)	0.10 (0.16)	237 (47.4)
		Low	0.12 (0.17)	282 (56.4)	0.13 (0.16)	273 (54.6)	0.13 (0.18)	278 (55.6)
	50%	Unif	0.27 (0.16)	419 (83.8)	0.25 (0.17)	399 (79.8)	0.25 (0.17)	402 (80.4)
		Mean	0.25 (0.17)	403 (80.6)	0.25 (0.17)	410 (82.0)	0.24 (0.17)	393 (78.6)
		Extreme	0.22 (0.16)	392 (78.4)	0.23 (0.17)	405 (81.0)	0.22 (0.16)	385 (77.0)
		High	0.23 (0.17)	398 (79.6)	0.23 (0.17)	384 (76.8)	0.25 (0.18)	398 (79.6)
		Low	0.24 (0.17)	408 (81.6)	0.25 (0.17)	403 (80.6)	0.27 (0.17)	415 (83.0)
	75%	Unif	0.37 (0.18)	475 (95.0)	0.38 (0.17)	479 (95.8)	0.37 (0.18)	467 (93.4)
		Mean	0.36 (0.17)	470 (94.0)	0.36 (0.16)	476 (95.2)	0.36 (0.16)	478 (95.6)
		Extreme	0.37 (0.17)	475 (95.0)	0.36 (0.17)	475 (95.0)	0.37 (0.17)	476 (95.2)
		High	0.38 (0.17)	477 (95.4)	0.36 (0.17)	468 (93.6)	0.38 (0.17)	474 (94.8)
		Low	0.38 (0.17)	477 (95.4)	0.38 (0.16)	475 (95.0)	0.39 (0.17)	481 (96.2)
8 items	25%	Unif	0.13 (0.20)	276 (55.2)	0.13 (0.20)	283 (56.6)	0.14 (0.21)	288 (57.6)
		Mean	0.13 (0.19)	279 (55.8)	0.14 (0.20)	285 (57.0)	0.12 (0.19)	274 (54.8)
		Extreme	0.11 (0.20)	248 (49.6)	0.10 (0.21)	242 (48.4)	0.10 (0.20)	255 (51.0)
		High	0.09 (0.20)	249 (49.8)	0.10 (0.20)	252 (50.4)	0.10 (0.19)	237 (47.4)
		Low	0.11 (0.19)	253 (50.6)	0.11 (0.21)	264 (52.8)	0.11 (0.20)	268 (53.6)
	50%	Unif	0.28 (0.20)	403 (80.6)	0.25 (0.20)	385 (77.0)	0.26 (0.20)	394 (78.8)
		Mean	0.24 (0.20)	367 (73.4)	0.23 (0.21)	364 (72.8)	0.24 (0.19)	372 (74.4)
		Extreme	0.22 (0.20)	355 (71.0)	0.22 (0.20)	363 (72.6)	0.20 (0.20)	353 (70.6)
		High	0.22 (0.20)	361 (72.2)	0.24 (0.19)	386 (77.2)	0.22 (0.19)	371 (74.2)
		Low	0.26 (0.20)	387 (77.4)	0.23 (0.19)	378 (75.6)	0.25 (0.20)	384 (76.8)
	75%	Unif	0.39 (0.20)	469 (93.8)	0.39 (0.20)	470 (94.0)	0.41 (0.19)	472 (94.4)
		Mean	0.37 (0.21)	455 (91.0)	0.36 (0.20)	449 (89.8)	0.36 (0.19)	454 (90.8)
		Extreme	0.35 (0.20)	452 (90.4)	0.32 (0.19)	433 (86.6)	0.34 (0.21)	439 (87.8)
		High	0.37 (0.19)	458 (91.6)	0.37 (0.19)	463 (92.6)	0.37 (0.20)	451 (90.2)
		Low	0.40 (0.20)	467 (93.4)	0.37 (0.20)	460 (92.0)	0.38 (0.20)	461 (92.2)

Note. Mean bias, standard deviation (*SD*) and frequency (%) of replications (over 500) in which bias ≥ 0.1 when DIF-size = 0.5, the group size = 100 and the differential item functioning (DIF) was uniform (γ: difference in the mean latent trait level between groups, *J*: number of items in the scale, *p*: % of DIF-items, *pos*: position of the DIF-items location parameters along the latent trait, Unif: uniformly distributed, Mean: around the mean of the items difficulties, Extreme: high and low difficulties, High: high difficulties, Low: low difficulties)

Appendix continues on the next page.

Appendix continues from the previous page.

Table 2C

Simulations with uniform DIF (DIF-size = 1, group size = 100)

			γ = 0		γ = 0.1		γ = 0.2	
J	*p*	*pos*	Bias Mean (*SD*)	Bias ≥ 0.1 *N* (%)	Bias Mean (*SD*)	Bias ≥ 0.1 *N* (%)	Bias Mean (*SD*)	Bias ≥ 0.1 *N* (%)
4 items	25%	Unif	0.24 (0.15)	407 (81.4)	0.24 (0.16)	406 (81.2)	0.23 (0.17)	388 (77.6)
		Mean	0.23 (0.16)	391 (78.2)	0.24 (0.15)	403 (80.6)	0.22 (0.16)	392 (78.4)
		Extreme	0.19 (0.16)	355 (71.0)	0.17 (0.16)	334 (66.8)	0.19 (0.17)	352 (70.4)
		High	0.19 (0.16)	368 (73.6)	0.18 (0.16)	339 (67.8)	0.18 (0.16)	349 (69.8)
		Low	0.24 (0.16)	408 (81.6)	0.24 (0.16)	411 (82.2)	0.23 (0.16)	398 (79.6)
	50%	Unif	0.48 (0.16)	496 (99.2)	0.49 (0.17)	495 (99.0)	0.46 (0.15)	496 (99.2)
		Mean	0.46 (0.15)	498 (99.6)	0.46 (0.16)	494 (98.8)	0.45 (0.16)	492 (98.4)
		Extreme	0.42 (0.16)	488 (97.6)	0.41 (0.15)	493 (98.6)	0.42 (0.15)	493 (98.6)
		High	0.42 (0.16)	490 (98.0)	0.43 (0.16)	490 (98.0)	0.42 (0.16)	493 (98.6)
		Low	0.49 (0.16)	498 (99.6)	0.48 (0.16)	499 (99.8)	0.47 (0.15)	497 (99.4)
	75%	Unif	0.70 (0.17)	500 (100.0)	0.70 (0.17)	500(100.0)	0.70 (0.18)	500 (100.0)
		Mean	0.68 (0.16)	500 (100.0)	0.66 (0.17)	500(100.0)	0.65 (0.17)	499 (99.8)
		Extreme	0.70 (0.17)	500 (100.0)	0.68 (0.16)	500(100.0)	0.68 (0.16)	500 (100.0)
		High	0.71 (0.17)	500 (100.0)	0.70 (0.17)	500(100.0)	0.70 (0.17)	500 (100.0)
		Low	0.75 (0.15)	500 (100.0)	0.73 (0.16)	500(100.0)	0.74 (0.16)	500 (100.0)
8 items	25%	Unif	0.24 (0.18)	394 (78.8)	0.26 (0.19)	402 (80.4)	0.25 (0.18)	392 (78.4)
		Mean	0.24 (0.18)	403 (80.6)	0.23 (0.17)	390 (78.0)	0.23 (0.17)	390 (78.0)
		Extreme	0.18 (0.20)	331 (66.2)	0.18 (0.19)	335 (67.0)	0.18 (0.19)	332 (66.4)
		High	0.14 (0.20)	289 (57.8)	0.15 (0.18)	330 (60.0)	0.16 (0.19)	307 (61.4)
		Low	0.24 (0.20)	376 (75.2)	0.23 (0.18)	377 (75.4)	0.22 (0.19)	366 (73.2)
	50%	Unif	0.49 (0.17)	496 (99.2)	0.51 (0.18)	491 (98.2)	0.47 (0.18)	486 (97.2)
		Mean	0.45 (0.18)	488 (97.6)	0.43 (0.18)	483 (96.6)	0.42 (0.18)	474 (94.8)
		Extreme	0.41 (0.18)	484 (96.8)	0.39 (0.18)	475 (95.0)	0.40 (0.19)	470 (94.0)
		High	0.43 (0.17)	490 (98.0)	0.42 (0.17)	483 (96.6)	0.43 (0.17)	490 (98.0)
		Low	0.50 (0.19)	493 (98.6)	0.48 (0.19)	484 (96.8)	0.48 (0.19)	492 (98.4)
	75%	Unif	0.75 (0.19)	500 (100.0)	0.74 (0.19)	500(100.0)	0.75 (0.19)	500 (100.0)
		Mean	0.69 (0.18)	500 (100.0)	0.69 (0.19)	497 (99.4)	0.69 (0.19)	500 (100.0)
		Extreme	0.66 (0.18)	500 (100.0)	0.65 (0.18)	499 (99.8)	0.64 (0.19)	497 (99.4)
		High	0.71 (0.19)	499 (99.8)	0.70 (0.18)	500(100.0)	0.70 (0.20)	499 (99.8)
		Low	0.74 (0.19)	500 (100.0)	0.74 (0.19)	500(100.0)	0.75 (0.20)	500 (100.0)

Note. Mean bias, standard deviation (*SD*) and frequency (%) of replications (over 500) in which bias ≥ 0.1 when DIF-size = 1, the group size = 100 and the differential item functioning (DIF) was uniform (γ: difference in the mean latent trait level between groups, *J*: number of items in the scale, *p*: % of DIF-items, *pos*: position of the DIF-items location parameters along the latent trait, Unif: uniformly distributed, Mean: around the mean of the items difficulties, Extreme: high and low difficulties, High: high difficulties, Low: low difficulties)

Appendix continues on the next page.

Appendix continues from the previous page.

Table 3B

Simulations with uniform DIF (DIF-size = 0.5, group size = 200)

			$\gamma = 0$		$\gamma = 0.1$		$\gamma = 0.2$	
			Bias	Bias ≥ 0.1	Bias	Bias ≥ 0.1	Bias	Bias ≥ 0.1
J	*p*	*pos*	Mean (*SD*)	*N* (%)	Mean (*SD*)	*N* (%)	Mean (*SD*)	*N* (%)
4 items	25%	Unif	0.13 (0.12)	284 (56.8)	0.13 (0.11)	305 (61.0)	0.13 (0.12)	294 (58.8)
		Mean	0.12 (0.11)	282 (56.4)	0.13 (0.12)	289 (57.8)	0.12 (0.12)	293 (58.6)
		Extreme	0.11 (0.12)	271 (54.2)	0.10 (0.11)	258 (51.6)	0.11 (0.12)	264 (52.8)
		High	0.11 (0.12)	266 (53.2)	0.11 (0.12)	256 (51.2)	0.10 (0.11)	255 (51.0)
		Low	0.13 (0.11)	298 (59.6)	0.11 (0.12)	272 (54.4)	0.11 (0.12)	271 (54.2)
	50%	Unif	0.26 (0.12)	462 (92.4)	0.26 (0.12)	459 (91.8)	0.26 (0.12)	449 (89.8)
		Mean	0.25 (0.12)	449 (89.8)	0.25 (0.12)	450 (90.0)	0.25 (0.12)	447 (89.4)
		Extreme	0.23 (0.12)	431 (86.2)	0.23 (0.12)	440 (88.0)	0.23 (0.12)	440 (88.0)
		High	0.24 (0.12)	441 (88.2)	0.24 (0.11)	452 (90.4)	0.23 (0.12)	423 (84.6)
		Low	0.25 (0.12)	454 (90.8)	0.25 (0.12)	454 (90.8)	0.25 (0.12)	447 (89.4)
	75%	Unif	0.37 (0.11)	496 (99.2)	0.37 (0.12)	497 (99.4)	0.38 (0.12)	499 (99.8)
		Mean	0.36 (0.12)	491 (98.2)	0.35 (0.12)	488 (97.6)	0.35 (0.12)	495 (99.0)
		Extreme	0.35 (0.12)	495 (99.0)	0.36 (0.12)	494 (98.8)	0.35 (0.12)	491 (98.2)
		High	0.37 (0.11)	497 (99.4)	0.37 (0.12)	491 (98.2)	0.37 (0.12)	492 (98.4)
		Low	0.38 (0.12)	492 (98.4)	0.38 (0.12)	500(100.0)	0.38 (0.12)	495 (99.0)
8 items	25%	Unif	0.14 (0.14)	299 (59.8)	0.13 (0.14)	280 (56.0)	0.13 (0.14)	303 (60.6)
		Mean	0.13 (0.14)	291 (58.2)	0.13 (0.14)	302 (60.4)	0.11 (0.13)	270 (54.0)
		Extreme	0.10 (0.14)	250 (50.0)	0.09 (0.14)	241 (48.2)	0.09 (0.13)	235 (47.0)
		High	0.09 (0.13)	233 (46.6)	0.10 (0.13)	240 (48.0)	0.09 (0.14)	230 (46.0)
		Low	0.11 (0.14)	264 (52.8)	0.12 (0.14)	261 (52.2)	0.09 (0.14)	239 (47.8)
	50%	Unif	0.26 (0.13)	447 (89.4)	0.27 (0.13)	451 (90.2)	0.26 (0.13)	444 (88.8)
		Mean	0.24 (0.14)	422 (84.4)	0.23 (0.13)	417 (83.4)	0.23 (0.14)	411 (82.2)
		Extreme	0.22 (0.13)	416 (83.2)	0.21 (0.14)	397 (79.4)	0.19 (0.14)	376 (75.2)
		High	0.23 (0.14)	419 (83.8)	0.23 (0.13)	422 (84.4)	0.22 (0.14)	405 (81.0)
		Low	0.26 (0.13)	443 (88.6)	0.26 (0.14)	436 (87.2)	0.24 (0.14)	419 (83.8)
	75%	Unif	0.39 (0.14)	490 (98.0)	0.39 (0.14)	490 (98.0)	0.38 (0.14)	492 (98.4)
		Mean	0.36 (0.14)	483 (96.6)	0.37 (0.13)	492 (98.4)	0.35 (0.14)	482 (96.4)
		Extreme	0.34 (0.14)	479 (95.8)	0.35 (0.14)	482 (96.4)	0.33 (0.14)	474 (94.8)
		High	0.37 (0.14)	491 (98.2)	0.37 (0.14)	485 (97.0)	0.37 (0.14)	487 (97.4)
		Low	0.39 (0.14)	489 (97.8)	0.37 (0.14)	485 (97.0)	0.37 (0.14)	490 (98.0)

Note. Mean bias, standard deviation (*SD*) and frequency (%) of replications (over 500) in which bias ≥ 0.1 when DIF-size = 0.5, the group size = 200 and the differential item functioning (DIF) was uniform (γ: difference in the mean latent trait level between groups, *J*: number of items in the scale, *p*: % of DIF-items, *pos*: position of the DIF-items location parameters along the latent trait, Unif: uniformly distributed, Mean: around the mean of the items difficulties, Extreme: high and low difficulties, High: high difficulties, Low: low difficulties)

Appendix continues on the next page.

Appendix continues from the previous page.

Table 3C

Simulations with uniform DIF (DIF-size = 1, group size = 200)

			γ = 0		γ = 0.1		γ = 0.2	
			Bias	Bias ≥ 0.1	Bias	Bias ≥ 0.1	Bias	Bias ≥ 0.1
J	*p*	*pos*	Mean (*SD*)	*N* (%)	Mean (*SD*)	*N* (%)	Mean (*SD*)	N (%)
4 items	25%	Unif	0.26 (0.11)	464 (92.8)	0.25 (0.12)	438 (87.6)	0.24 (0.12)	447 (89.4)
		Mean	0.23 (0.11)	438 (87.6)	0.22 (0.11)	433 (87.6)	0.22 (0.11)	435 (87.0)
		Extreme	0.19 (0.11)	396 (79.2)	0.17 (0.11)	372 (74.4)	0.17 (0.11)	368 (73.6)
		High	0.18 (0.12)	377 (75.4)	0.18 (0.11)	388 (77.6)	0.18 (0.11)	378 (75.6)
		Low	0.24 (0.11)	448 (89.6)	0.24 (0.11)	443 (88.6)	0.23 (0.11)	448 (89.6)
	50%	Unif	0.48 (0.12)	500 (100.0)	0.47 (0.11)	500(100.0)	0.47 (0.11)	500 (100.0)
		Mean	0.46 (0.11)	500 (100.0)	0.45 (0.11)	500(100.0)	0.45 (0.11)	500 (100.0)
		Extreme	0.41 (0.10)	500 (100.0)	0.40 (0.11)	499 (99.8)	0.40 (0.11)	497 (99.4)
		High	0.42 (0.11)	500 (100.0)	0.42 (0.12)	500(100.0)	0.41 (0.11)	500 (100.0)
		Low	0.48 (0.11)	500 (100.0)	0.47 (0.11)	500(100.0)	0.48 (0.11)	500 (100.0)
	75%	Unif	0.69 (0.11)	500 (100.0)	0.69 (0.12)	500(100.0)	0.69 (0.12)	500 (100.0)
		Mean	0.67 (0.11)	500 (100.0)	0.67 (0.12)	500(100.0)	0.65 (0.11)	500 (100.0)
		Extreme	0.69 (0.11)	500 (100.0)	0.68 (0.11)	500(100.0)	0.67 (0.12)	500 (100.0)
		High	0.71 (0.12)	500 (100.0)	0.69 (0.12)	500(100.0)	0.69 (0.12)	500 (100.0)
		Low	0.73 (0.11)	500 (100.0)	0.74 (0.12)	500(100.0)	0.72 (0.12)	500 (100.0)
8 items	25%	Unif	0.25 (0.13)	433 (86.6)	0.25 (0.13)	446 (89.2)	0.24 (0.12)	429 (85.8)
		Mean	0.24 (0.13)	432 (86.4)	0.24 (0.12)	433 (86.6)	0.23 (0.13)	420 (84.0)
		Extreme	0.18 (0.13)	369 (73.8)	0.17 (0.13)	351 (70.2)	0.18 (0.13)	354 (70.8)
		High	0.16 (0.13)	345 (69.0)	0.15 (0.12)	313 (62.6)	0.15 (0.13)	344 (68.8)
		Low	0.23 (0.14)	403 (80.6)	0.23 (0.13)	407 (81.4)	0.21 (0.13)	404 (80.8)
	50%	Unif	0.49 (0.12)	500 (100.0)	0.48 (0.13)	500(100.0)	0.47 (0.12)	499 (99.8)
		Mean	0.43 (0.13)	496 (99.2)	0.43 (0.12)	497 (99.4)	0.42 (0.12)	495 (99.0)
		Extreme	0.39 (0.13)	492 (98.4)	0.39 (0.13)	492 (98.4)	0.39 (0.13)	494 (98.8)
		High	0.42 (0.12)	497 (99.4)	0.42 (0.12)	499 (99.8)	0.40 (0.12)	497 (99.4)
		Low	0.48 (0.13)	500 (100.0)	0.48 (0.13)	500(100.0)	0.46 (0.13)	496 (99.2)
	75%	Unif	0.75 (0.13)	500 (100.0)	0.74 (0.13)	500(100.0)	0.72 (0.13)	500 (100.0)
		Mean	0.69 (0.13)	500 (100.0)	0.69 (0.13)	500(100.0)	0.68 (0.13)	500 (100.0)
		Extreme	0.65 (0.13)	500 (100.0)	0.64 (0.13)	500(100.0)	0.64 (0.13)	500 (100.0)
		High	0.70 (0.13)	500 (100.0)	0.70 (0.13)	500(100.0)	0.69 (0.13)	500 (100.0)
		Low	0.74 (0.14)	500 (100.0)	0.75 (0.13)	500(100.0)	0.72 (0.13)	500 (100.0)

Note. Mean bias, standard deviation (*SD*) and frequency (%) of replications (over 500) in which bias ≥ 0.1 when DIF-size = 1, the group size = 200 and the differential item functioning (DIF) was uniform (γ: difference in the mean latent trait level between groups, *J*: number of items in the scale, *p*: % of DIF-items, *pos*: position of the DIF-items location parameters along the latent trait, Unif: uniformly distributed, Mean: around the mean of the items difficulties, Extreme: high and low difficulties, High: high difficulties, Low: low difficulties)