Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression

Scott, N.W.; Fayers, P.M.; Aaronson, N.K.; Bottomley, A.; de Graeff, A.; Groenvold, M.; Gundy, C.; Koller, M.; Petersen, M.A.; Sprangers, M.A.G.

DOI
10.1186/1477-7525-8-81

Publication date
2010

Document Version
Final published version

Published in
Health and Quality of Life Outcomes

HEALTH AND QUALITY OF LIFE OUTCOMES

RESEARCH Open Access

Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression

Neil W Scott[1*], Peter M Fayers[1,2], Neil K Aaronson[3], Andrew Bottomley[4], Alexander de Graeff[5], Mogens Groenvold[6,7], Chad Gundy[3], Michael Koller[8], Morten A Petersen[6], Mirjam AG Sprangers[9], the EORTC Quality of Life Group and the Quality of Life Cross-Cultural Meta-Analysis Group

Abstract

Background: Differential item functioning (DIF) methods can be used to determine whether different subgroups respond differently to particular items within a health-related quality of life (HRQoL) subscale, after allowing for overall subgroup differences in that scale. This article reviews issues that arise when testing for DIF in HRQoL instruments. We focus on logistic regression methods, which are often used because of their efficiency, simplicity and ease of application.

Methods: A review of logistic regression DIF analyses in HRQoL was undertaken. Methodological articles from other fields and using other DIF methods were also included if considered relevant.

Results: There are many competing approaches for the conduct of DIF analyses and many criteria for determining what constitutes significant DIF. DIF in short scales, as commonly found in HRQL instruments, may be more difficult to interpret. Qualitative methods may aid interpretation of such DIF analyses.

Conclusions: A number of methodological choices must be made when applying logistic regression for DIF analyses, and many of these affect the results. We provide recommendations based on reviewing the current evidence. Although the focus is on logistic regression, many of our results should be applicable to DIF analyses in general. There is a need for more empirical and theoretical work in this area.

Background

Many health-related quality of life (HRQoL) instruments contain multi-item scales. As part of the process of validating a HRQoL instrument it may be desirable to know whether each item behaves in the same way for different subgroups of respondents. For example, do males and females respond differently to a question about carrying heavy objects, even after accounting for their overall level of physical functioning? Is an item about fatigue answered similarly by older and younger age groups, given the same overall fatigue level? Does a translation of a questionnaire item behave in the same way as the original version? Differential item functioning (DIF) methods are a range of techniques that are increasingly being used to evaluate whether different subgroups respond differently to particular items within a scale, after controlling for group differences in the overall HRQoL domain being assessed.

DIF analyses were first used in educational testing settings to investigate whether particular items in a test were unfair to, for example, females or a particular ethnic group, even after adjusting for that group's overall test ability. In HRQoL research, similar analyses may be used to assess whether there are differences in response to a particular subscale item as a function of respondent characteristics such as age group, gender, education or treatment, given the same level of HRQoL. DIF analyses may also be employed to evaluate cross-cultural response differences, e.g. by country or ethnicity or to evaluate translations of questionnaire items. Whereas in educational settings, items with DIF may simply be

* Correspondence: n.w.scott@abdn.ac.uk
[1]Section of Population Health, University of Aberdeen, UK
Full list of author information is available at the end of the article

dropped or replaced, this may be less straightforward in HRQoL settings if an instrument is already established.

DIF analyses can be carried out using a wide range of statistical methods to explore the relationship between three variables: is group membership (g) associated with differential responses (x_i) to an item (x) for respondents at the same level of a matching criterion (θ)? For example, DIF analyses examining the effect of gender on a particular pain item consider not only the proportions of males and females choosing each item category, but also the possibility that males and females report different levels of overall pain as measured by the other pain items.

The **grouping variable** (or **exogenous variable**) g may be binary, such as male/female, or may have multiple categories. The item response (x_i) may be binary (e.g. yes/no) or ordered categorical (e.g. good/fair/poor). The **matching criterion** or **matching variable** (θ) is used to account for different levels of functioning or ability in each group. For some DIF methods, an observed scale score (frequently the sum of the items) is used as the matching variable; in other methods a latent variable is used.

Two distinct types of DIF can be distinguished. **Uniform DIF** occurs if an item shows the same amount of DIF whatever the level of θ. When **non-uniform DIF** is present, the magnitude of the effect varies according to θ. For example, non-uniform gender DIF might occur in a pain item if it were found that males with lower levels of pain were more likely to score higher on an item compared with female respondents, whereas males with severe pain might be relatively less likely than females to score highly. Detection procedures should attempt to assess both uniform and non-uniform DIF, although in practice not all methods can detect non-uniform DIF.

The literature on DIF is diverse because there is a wide choice of methodologies that may be employed, including contingency table, item response theory (IRT), structural equation modelling and logistic regression methods. Although these represent very different methodological approaches, there are also many challenges that may be encountered regardless of the DIF method used. One widely used approach for detecting DIF is logistic regression, which is commonly regarded as simple, robust and reasonably efficient, while being easy to implement. This paper focuses primarily on the use of the logistic regression method, although many of the conclusions are likely to be equally pertinent to other DIF methods, and is intended to complement existing review articles on logistic regression DIF [1,2], which have a somewhat different focus to our review.

Aim

The specific aim of this article is to provide an overview of the logistic regression approach to DIF detection. The review also considers more general methodological issues specific to DIF analyses of HRQoL instruments, including the evaluation of DIF in short scales and the problems with interpreting DIF.

Methods

Although this should not be considered a systematic review as judgement was used to select included articles, a systematic search strategy using the search term "differential item functioning" was employed to identify relevant articles using the electronic databases MEDLINE, EMBASE and Web of Knowledge. Abstracts of the articles were assessed for relevance and a decision made whether or not to review the full article. Priority was given to studies concerning HRQoL instruments, but as DIF analyses originated in educational testing, much of the literature relates to educational settings. DIF studies from other areas were therefore included if considered to have broader methodological relevance. Although the greatest emphasis was placed on articles using logistic regression techniques, articles relating to any DIF methodology were included if considered relevant to the discussion of specific issues or topics. The electronic literature search was supplemented by relevant articles and books from the reference lists of studies already included.

Results

A total of 211 (MEDLINE), 147 (EMBASE) and 589 (Web of Knowledge) articles met the initial search criteria. The full text of 136 articles was accessed as part of the review.

DIF detection studies were identified for HRQoL instruments from many clinical areas including: asthma [3], oncology [4-9], headache [10,11], mental health [12-18] and functional ability [19-21].

A wide range of grouping factors has been evaluated in HRQoL DIF studies including: language/translation [7,8,11,12,22], language group [23], country [5,16,19,21,22,24,25], gender [3,10,13,14,17,19,22,25-30], age [4,10,22,25,27,29,30], ethnicity [6,13,15,27,29-31], education [10,28,29], employment status [10], job category [32], treatment [4] and type of condition [22,20].

Methods for Investigating DIF

A large number of diverse statistical methods for detecting DIF have been described in the literature [33-38]. DIF methods may be divided into parametric methods, requiring distributional assumptions of a particular model, and non-parametric methods that are distribution-free. Provided that the assumptions are met, parametric approaches may be more powerful and stable [37].

Many DIF detection studies have used methods based on item response theory (IRT) [35,39], including a

number of recent studies of HRQoL instruments [5,6,20,40]. The main advantage of IRT DIF techniques is the use of a latent (rather than an observed) variable for θ, the matching criterion. Disadvantages include possible lack of model fit, increased sample size requirements and the need for more specialised computer software [41].

Contingency table methods, particularly the Mantel-Haenszel and standardisation approaches, are non-parametric methods that are frequently used in educational testing [42,43]. These methods are straightforward to perform and do not require any model assumptions to be satisfied, but are unable to detect non-uniform DIF. These methods have been infrequently used in HRQoL research, although an approach using the partial gamma statistic has been used [36]. Other DIF detection methods include the simultaneous item bias test (SIBTEST) method [44] and approaches using structural equation modelling [45].

Logistic regression

The remainder of this review will concentrate on the method of logistic regression [1,2,46-49].

For items with two response categories, binary logistic regression can be used to relate the probability of positive response (p) to the grouping variable (g), the total scale score (representing ability level/level of quality of life) (θ) and the interaction of the group and scale score (the product of g and θ). In HRQoL research, items frequently have three or more ordered response categories, necessitating use of ordinal logistic regression instead. This estimates a single common odds ratio assuming that the odds are proportional across all categories [50].

The binary and ordinal logistic regression models can be written respectively as:

$$\ln\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1\theta + \beta_2 g + \beta_3 g\theta$$

$$\ln\left[\frac{\Pr(Y \le k \mid g,\theta)}{1-\Pr(Y \le k \mid g,\theta)}\right] = \beta_{0k} + \beta_1\theta + \beta_2 g + \beta_3 g\theta \ \ (k = 0,1,2,\ldots)$$

where $\Pr(Y \le k)$ is the probability of response in category k or below (k = 0,1,2,...) and $\beta_{0k}, \beta_1, \beta_2, \beta_3$ are constants usually estimated by maximum likelihood.

An advantage of logistic regression methods is the ability to test for both uniform and non-uniform DIF. The presence of uniform DIF is evaluated by testing whether the regression coefficient of group membership (β_2) differs significantly from zero. A test of the interaction coefficient between group membership and ability (β_3) can be used to assess non-uniform DIF.

Some authors advocate first testing the presence of both uniform and non-uniform DIF simultaneously using a test of the null hypothesis that $\beta_2 = \beta_3 = 0$ [2,46,47]. The difference in the -2 Log Likelihood (-2LL) of these models is assessed using a chi-squared distribution with two degrees of freedom (2 df). If this step gives a significant result, the presence of uniform DIF alone is then determined by testing the significance of β_2 using a chi-squared distribution with one degree of freedom (1 df). An alternative strategy is to report two separate 1 df chi-squared tests for uniform and non-uniform DIF [51]. Simulations have shown that this approach may lead to improved performance [49,52].

Perhaps the main advantage of the logistic regression DIF approach is its flexibility [2,53]. For example, if more than two groups are to be compared, extra variables may be included in the regression model to indicate the effect of each group with respect to a reference category. Another advantage is the ease of adjusting for additional covariates, both continuous and categorical, which may confound the DIF analyses. Despite this much-cited benefit, few logistic regression DIF studies making use of adjusted analyses were identified [8]. In fact, given interpretation difficulties, some authors prefer to test each covariate for DIF in separate models [54].

Methodological issues with DIF Analyses

Sample size

There are no established guidelines on the sample size required for DIF analyses. The minimum number of respondents will depend on the type of method used, the distribution of the item responses in the two groups, and whether there are equal numbers in each group. For binary logistic regression it has been found that 200 per group is adequate [1], and a sample size of 100 per group has also been reported to be acceptable for items without skewness [55]. For ordinal logistic regression, simulations suggested that 200 per group may be adequate, except for two-item scales [56]. As a general rule of thumb, we suggest a minimum of 200 respondents per group as a requirement for logistic regression DIF analyses.

Unidimensionality

DIF analyses assume that the underlying distribution of θ is unidimensional [34], with all items measuring a single concept; in fact, some authors suggest that DIF is itself a form of multidimensionality [38]. Although it has been recommended that factor analysis methods be used to confirm unidimensionality prior to performing DIF analyses [38], in practice few DIF studies have reported dimensionality analyses [57]. When the construct validation of a HRQoL instrument has already explored scale dimensionality, further testing may be deemed unnecessary.

Deriving the matching criterion

It might seem counter-intuitive to include the studied item itself when calculating a scale score for the

matching criterion, but studies have found that DIF detection was more accurate when this is done [35,58]. Thus, if the matching criterion is the summated scale score, the item being studied should not be excluded from the summation.

Purification

An item with DIF might bias the scale score estimate, making it less valid as a matching criterion for other items. Some DIF studies have employed "purification" [35], which is an iterative process of eliminating items with the most severe DIF from the matching criterion when assessing other items. Purification has been shown to be beneficial in DIF analyses in other fields [59,60], but has rarely been used in HRQoL research [61], perhaps owing to the lower number of items in HRQoL subscales. We recommend that more consideration be given to purification, although the benefit may depend on the number of items in the scale: it may be less suitable for scales with just a small number of items, as removing items can affect the precision of the matching variable. For these scales, we would recommend more qualitative approaches that attempt to understand underlying reasons for DIF.

Sum scoring versus IRT scoring

An important disadvantage of the logistic regression method is reliance on an observed scale score, which may not be an adequate matching variable, particularly for short scales [53,62]. Thus, it has been suggested that item response theory (IRT) scoring should be used to derive the matching variable, even when IRT is not itself used for DIF detection. This hybrid logistic regression/IRT method has been used in a number of recent studies and free software is available for this purpose [2,62,63]. It also has the advantage of incorporating purification by using an iterative approach that can account for DIF in other items [63,64]. It is our view, however, that the standard logistic regression approach using sum scores is an acceptable method in practice; reported results of DIF analyses using the hybrid method have tended to be similar to those obtained using sum scores [2].

Pseudo-DIF

"Pseudo-DIF" results when DIF in one item causes apparent opposing DIF in other items in the same scale, even though these other items are not biased [36]. For example, in logistic regression DIF analyses the log odds ratios for items in a scale will sum approximately to zero. Thus log odds ratios for items without real DIF may be forced into the opposite direction to compensate for items with true DIF. The most extreme case occurs for two-item scales where opposite DIF effects will be found for the two items; the results are therefore impossible to interpret without additional external information (see the section on qualitative methods below) [65].

Scale length and floor/ceiling effects

In HRQoL research the number of items per scale may vary, and subscales may often contain only a few items in order to minimise the burden on patients. DIF analyses of short scales may be difficult to interpret because of pseudo-DIF and the scale score may also be a less accurate measure of the underlying construct. Several studies have successfully conducted DIF analyses in scales with fewer than ten items [3-5,7-9,11,19,20,22,24,61].

Another common problem with HRQoL instruments is items with floor and ceiling effects, or with highly skewed score distributions. These items will not be able to discriminate between groups as effectively as other items [35,37]. Simulations show that there is reduced power to detect DIF in such items, although Type I error rates appear to be stable [56].

Interpretation of DIF Analyses

Like many other DIF detection methods, logistic regression uses statistical hypothesis tests to identify DIF. Interpretation of an item with statistically significant DIF is rarely straightforward. It could have arisen purely by chance, it could result from pseudo-DIF in another item in the same scale, or it could be caused by confounding [7,36]. If real DIF does exist there might be more than one possible cause. For example, for DIF analyses of a questionnaire with respect to country, observed DIF could either be caused by a lack of translation equivalence or by cross-cultural response differences. Sample size also affects interpretation of DIF - sufficiently large sample sizes may result in the detection of unimportant yet statistically significant DIF.

Methods of adjustment for multiple testing

Multiple hypothesis testing may be a particular problem in DIF analyses: there may be more than one HRQoL subscale of interest, analyses may be performed for all items within the scales, and for each item there may be several grouping variables. If some of these grouping variables have several categories (e.g. the translation used), this may involve several tests for each variable. Finally, tests for both uniform and non-uniform DIF may be conducted. The large number of significance tests increases the probability of obtaining false statistically significant results by chance alone.

Multiple testing is common to many statistical applications and the various approaches to address these issues are reviewed elsewhere [66]. One solution is to use a Bonferroni approach (dividing the nominal

statistical significance level, typically 0.05, by the number of tests conducted); this reduces the Type I errors, but is a very conservative approach. Some DIF studies have used a 1% significance level instead [19,55,67]. An alternative approach is to use cross-validation, whereby the data are randomly divided into two datasets, and one of the halves is used to confirm the results obtained on the other half [4,24]. In general, researchers investigating DIF should account for the number of significance tests conducted, unless they regard the search for DIF as hypothesis-generating and report their findings as tentative, in which case multiple testing is arguably less of an issue [62].

Methods of determining clinical significance

Since statistical significance does not necessarily imply clinical or practical significance, many authors have proposed DIF classifications that incorporate both statistical significance and the magnitude of DIF, but once again the question of which thresholds to use is not straightforward.

One widely used approach is first to calculate statistical significance using the standard likelihood ratio test and then to calculate, as a measure of effect size, the change in the R^2 associated with including the grouping variable in the model. For ordinal logistic regression a measure such as McKelvey and Zavoina's pseudo-R^2 may be used [1]. Non-uniform DIF may be assessed similarly [68].

Two sets of rules have been developed to classify DIF using the change in R^2, the Zumbo-Thomas procedure [1] and the Jodoin-Gierl approach [49]. The corresponding cut-offs for indicating moderate and large DIF are very different: 0.13 and 0.26 for Zumbo-Thomas and 0.035 and 0.070 for Jodoin-Gierl. Both systems usually require a p-value of less than 0.001. Unsurprisingly, these criteria can produce very different numbers of items flagged with DIF [49,69] and several authors have also remarked that Zumbo's method is very conservative and that few items meet the criteria [23,55]. An R^2 difference cut-off level of 0.02 has also been suggested by Bjorner et al. (2003), and used in other studies [10,11,22,25], whereas Kristensen et al. (2004) used a rule that the group variable had to explain at least 5% of the item variation after adjusting for the sum score [32].

Crane has suggested testing for non-uniform DIF using a Bonferroni-corrected likelihood ratio chi-squared test with 1 df. For uniform DIF, significance criteria are not used: the change in the regression coefficient for θ in models with and without the group variable is calculated and a 10% difference is used to indicate important DIF [2,62]. In a more recent study, a 5% difference was used [63].

In logistic regression DIF analyses, the odds ratio associated with the grouping variable can also be used as a magnitude criterion. For example, Cole et al. (2000) used proportional odds ratios greater than 2 or less than 0.5 to denote practically meaningful DIF [27]. A classification system adapted from that used in educational testing has also been used with odds ratios [70]. Slight to moderate DIF is indicated by a statistically significant odds ratio that is also outside the interval 0.65 to 1.53; moderate to large DIF is indicated if the odds ratio is outside 0.53 to 1.89 and significantly less than 0.65 or greater than 1.53 [24]. A number of studies have used a threshold in the log odds ratios of 0.64 (≈ln(0.53)), often in conjunction with $p < 0.001$ [7-9,61].

A recent study compared three assessment criteria for evaluating two composite scales formed from items taken from a number of HRQoL instruments [71]: Swaminathan and Roger's approach using only statistical significance [46], Zumbo and Gelin's pseudo-R^2 magnitude criterion [14], and Crane's 5% change in the regression coefficient [2]. The three methods flagged very different numbers of items as having DIF. This is not surprising and stems partly from the dichotomisation of DIF effects into either DIF or no DIF, when in fact it is a matter of degree [72]. There is currently no consensus regarding effect size classification system for logistic regression DIF analyses, and there is a need for further investigation [49]. What is of primary importance is that results of the statistical significance tests should not be interpreted without reference to their clinical significance.

Illustration of DIF

Some authors advocate the use of graphical methods to display the magnitude and direction of DIF effects [73]. Forest plots may provide a convenient way to summarise the pattern of DIF across several categories [8]. Crane's logistic regression software produces box and whisker plots to evaluate the impact of DIF on each covariate [63,74,75].

What should be done if DIF is found?

Unfortunately, the DIF literature tends to focus on how to detect DIF, rather than on what to do when it is found, but there are two main steps that may be employed. First, if significant DIF, uniform or non-uniform, is found, detailed examination of the three-way contingency table of item, scale score and grouping variable can help interpret the direction and nature of this DIF effect. It may then be helpful to identify underlying reasons for the differential functioning using expert item review (see the section on qualitative methods below).

The second approach is to determine the practical impact of observed DIF. This can be assessed, for

example, by removing items with DIF and determining what difference this makes to the results [76]. Impact analyses have also been used to investigate whether item-level DIF results in clinically important differences at the scale level [77]. Some authors have attempted to use IRT methods to adjust their results and correct for the presence of DIF [6,7,63]. Others have argued that at the scale level DIF due to multidimensionality may in fact balance out [78].

If an instrument is at the development stage, modifications can also be made to items before retesting in further DIF analyses. If translation DIF is found for a particular item, the wording may be reviewed by independent translators. It becomes more problematic when a DIF effect is found for an established HRQoL questionnaire: researchers need to consider carefully how this will affect future studies. For example, if DIF is found with respect to age group, this may not be important for a study with narrow age inclusion criteria, but it would be for studies including both older and younger participants. DIF may also have lower impact on clinical trials than on observational studies as randomisation may ensure groups are balanced with respect to important patient characteristics [77].

Use of qualitative methods alongside DIF analyses

Some authors have attempted to interpret the underlying causes of flagged DIF, either anecdotally or by using formal qualitative methods. Studies in the educational field have, however, typically found low agreement between expert reviews of items and statistical DIF analyses [34,57]. For example, many HRQoL instruments are translated into other languages or undergo cultural adaptation for use in another country. DIF analyses may be useful for evaluating item translations and, if DIF is found, the relevant wording may be reviewed. It may be difficult, however, to separate lack of translation equivalence from cross-cultural response differences.

We identified only a few studies that attempted to relate DIF results to blinded substantive assessments of the reasons for DIF: most conducted in fields such as educational testing [8,67,79-85]. A number of studies attempted to give post hoc explanations for DIF effects found in HRQoL instruments [4,6,7,12,16,19,22, 24,25,86]. Where resources exist to do this, we recommend that researchers employ expert review of DIF items as part of the process of understanding and interpreting DIF effects. They are particularly useful in situations with more than one possible source of DIF, such as when distinguishing between cultural and linguistic response differences in DIF analyses of translations. A more detailed review of the studies using external information alongside DIF analyses may be found elsewhere [65].

Summary

Although much of the published research on DIF methods concerns educational tests, DIF techniques are increasingly being applied to HRQoL outcomes. This introduces a new set of challenges. HRQoL scales often consist of short scales with ordered categorical items, and some items may exhibit floor and ceiling effects. Pseudo-DIF may be a problem, and without parallel qualitative methods the underlying causes of the DIF effects may not be clear.

Many methods for DIF detection are available, and this review has focused largely on just one such approach: logistic regression. This method has several advantages in the context of HRQoL DIF analyses, but a disadvantage is the reliance on sum scores as the matching variable. IRT DIF methods using a latent matching variable have important theoretical advantages but these may be less accessible to those with only standard statistical software. The hybrid logistic regression/IRT method has been employed successfully in several studies although the evidence of tangible practical benefit over the standard sum score method is limited.

There are many competing criteria for determining what constitutes important DIF, using either statistical significance or magnitude criteria, and these have been shown to flag different numbers of items with DIF. In educational contexts the level of DIF that is important is a matter of policy, and practical considerations are most important [35]. Similarly, although DIF analysis is an important tool in HRQoL research, it cannot be employed on its own: judgement should be used alongside the statistical results when deciding whether a particular DIF effect is of sufficient practical importance to require modification of an item or scale.

The choices made during analysis will substantially affect the results, and we have described and illustrated the impact of these choices. We have reviewed the literature and provided guidance for making the decisions about the optimal application of logistic regression for DIF analysis. Many of these findings are likely to be equally pertinent to other approaches for detecting DIF.

Key Messages

- A variety of DIF methodologies are available. For HRQoL instruments, logistic regression is a robust and flexible method and therefore a good practical choice in most situations. A hybrid logistic regression/IRT method, which avoids the theoretical disadvantages of using the sum score as a matching variable, is also available.
- A combination of statistical significance and magnitude criteria should be used when classifying items as having DIF. When interpreting results, allowance should be made for the number of tests conducted.

- When deriving the matching criterion for logistic regression DIF using sum scores, the overall scale score including the studied item should be used.
- For longer scales researchers should consider iteratively eliminating items with DIF in subsequent DIF analyses (purification).
- Prior to conducting DIF analyses, it should be checked that a scale is unidimensional.
- At least 200 respondents per group are recommended for logistic regression DIF analyses.
- Graphical methods may be used to display DIF results in multiple groups.

Acknowledgements of research support
This work was funded by the European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Group, Cancer Research UK and the University of Aberdeen and carried out under the auspices of the EORTC Quality of Life Group.

Author details
[1]Section of Population Health, University of Aberdeen, UK. [2]Department of Cancer Research and Molecular Medicine, Faculty of Medicine, Norwegian University of Science and Technology, Trondheim, Norway. [3]Division of Psychosocial Research and Epidemiology, Netherlands Cancer Institute, Amsterdam, Netherlands. [4]Quality of Life Department, European Organisation for Research and Treatment of Cancer Headquarters, Brussels, Belgium. [5]Division of Medical Oncology, Department of Internal Medicine, University Medical Centre, Utrecht, Netherlands. [6]Department of Palliative Medicine, Bispebjerg Hospital, Copenhagen, Denmark. [7]Institute of Public Health, University of Copenhagen, Denmark. [8]Centre for Clinical Studies, University Hospital Regensburg, Regensburg, Germany. [9]Department of Medical Psychology, Academic Medical Centre, University of Amsterdam, Netherlands.

Authors' contributions
NWS conducted the literature review and wrote the first draft of the article. PMF, NKA, AB, AdG, MG, CG, MK, MAP and MAGS contributed to subsequent drafts. All authors read and approved the final version.

Competing interests
The authors declare that they have no competing interests.

Received: 17 December 2009 Accepted: 4 August 2010
Published: 4 August 2010

References

1. Zumbo BD: **A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.** Ottowa, ON: Directorate of Human Research and Evaluation, Department of National Defense 1999.
2. Crane PK, Gibbons LE, Jolley L, van Belle G: **Differential item functioning analysis with ordinal logistic regression techniques: DIFdetect and difwithpar.** *Med Care* 2006, **44**:S115-S123.
3. Gelin MN, Carleton BC, Smith MA, Zumbo BD: **The dimensionality and gender differential item functioning of the mini asthma quality of life questionnaire (MINIAQLQ).** *Soc Indicators Res* 2004, **68**:91-105.
4. Groenvold M, Bjorner JB, Klee MC, Kreiner S: **Test for item bias in a quality of life questionnaire.** *J Clin Epidemiol* 1995, **48**:805-816.
5. Hahn EA, Holzner B, Kemmler G, Sperner-Unterweger B, Hudgens SA, Cella D: **Cross-cultural evaluation of health status using item response theory: FACT-B comparisons between Austrian and U.S. patients with breast cancer.** *Eval Health Prof* 2005, **28**:233-259.
6. Pagano IS, Gotay CC: **Ethnic differential item functioning in the assessment of quality of life in cancer patients.** *Health and Quality of Life Outcomes* 2005, **3**:1-10.
7. Petersen MA, Groenvold M, Bjorner JB, Aaronson N, Conroy T, Cull A, Fayers P, Hjermstad M, Sprangers M, Sullivan M, European Organisation for Research and Treatment of Cancer Quality of Life, Group: **Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire.** *Quality of Life Research* 2003, **12**:373-385.
8. Scott NW, Fayers PM, Bottomley A, Aaronson NK, de Graeff A, Groenvold M, Koller M, Petersen MA, Sprangers MAG: **Comparing translations of the EORTC QLQ-C30 using differential item functioning analyses.** *Quality of Life Research* 2006, **15**:1103-1115.
9. Scott NW, Fayers PM, Aaronson NK, Bottomley A, de Graeff A, Groenvold M, Koller M, Petersen MA, Sprangers MAG: **The use of differential item functioning analyses to identify cultural differences in responses to the EORTC QLQ-C30.** *Quality of Life Research* 2007, **16**:115-129.
10. Bjorner JB, Kosinski M, Ware JE: **Calibration of an item pool for assessing the burden of headaches: An application of item response theory to the headache impact test (HIT).** *Quality of Life Research* 2003, **12**:913-933.
11. Martin M, Blaisdell B, Kwong JW, Bjorner JB: **The short-form headache impact test (HIT-6) was psychometrically equivalent in nine languages.** *J Clin Epidemiol* 2004, **57**:1271-1278.
12. Azocar F, Arean P, Miranda J, Munoz RF: **Differential item functioning in a Spanish translation of the Beck Depression Inventory.** *J Clin Psychol* 2001, **57**:355-365.
13. Dancer LS, Anderson AJ, Derlin RL: **Use of log-linear models for assessing differential item functioning in a measure of psychological functioning.** *Journal of Consulting & Clinical Psychology* 1994, **62**:710-717.
14. Gelin MN, Zumbo BD: **Differential item functioning results may change depending on how an item is scored: An illustration with the Center for Epidemiologic Studies Depression Scale.** *Educational and Psychological Measurement* 2003, **63**:65-74.
15. Iwata N, Turner RJ, Lloyd DA: **Race/ethnicity and depressive symptoms in community-dwelling young adults: A differential item functioning analysis.** *Psychiatry Res* 2002, **110**:281-289.
16. Iwata N, Buka S: **Race/ethnicity and depressive symptoms: A cross-cultural/ethnic comparison among university students in East Asia, North and South America.** *Soc Sci Med* 2002, **55**:2243-2252.
17. Zumbo BD, Gelin MN, Hubley AM: **Psychometric study of the CES-D: Factor analysis and DIF.** Presented at the International Neuropsychological Society Annual Meeting, Chicago 2001.
18. Orlando M, Marshall GN: **Differential item functioning in a Spanish translation of the PTSD checklist: Detection and evaluation of impact.** *Psychol Assess* 2002, **14**:50-59.
19. Avlund K, Era P, Davidsen M, GauseNilsson I: **Item bias in self-reported functional ability among 75-year-old men and women in three Nordic localities.** *Scand J Soc Med* 1996, **24**:206-217.
20. Dallmeijer AJ, Dekker J, Roorda LD, Knol DL, van Baalen B, de Groot V, Schepers VPM, Lankhorst GJ: **Differential item functioning of the functional independence measure in higher performing neurological patients.** *J Rehabil Med* 2005, **37**:346-352.
21. Tennant A, Penta M, Tesio L, Grimby G, Thonnard JL, Slade A, Lawton G, Simone A, Carter J, Lundgren-Nilsson A, Tripolski M, Ring H, Biering-Sorensen F, Marincek C, Burger H, Phillips S: **Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: The PRO-ESOR project.** *Med Care* 2004, **42**:37-48.
22. Schmidt S, Mühlan H, Power M: **The EUROHIS-QOL 8-item index: Psychometric results of a cross-cultural field study.** *European Journal of Public Health Advance Access* 2006, **16**:420-428.
23. Kim M: **Detecting DIF across the different language groups in a speaking test.** *Language Testing* 2001, **18**:89-114.
24. Bjorner JB, Kreiner S, Ware JE, Damsgaard MT, Bech P: **Differential item functioning in the Danish translation of the SF-36.** *J Clin Epidemiol* 1998, **51**:1189-1202.
25. Schmidt S, Debensason D, Mühlan H, Petersen C, Power M, Simeoni MC, Bullinger M: **The DISABKIDS generic quality of life instrument showed cross-cultural validity.** *Journal of Clinical Epidemiology* 2006, **59**:587-598.
26. Borsboom D, Mellenbergh GJ, van Heerden J: **Different kinds of DIF: A distinction between absolute and relative forms of measurement invariance and bias.** *Applied Psychological Measurement* 2002, **26**:433-450.
27. Cole SR, Kawachi I, Maller SJ, Berkman LF: **Test of item-response bias in the CES-D scale. Experience from the New Haven EPESE study.** *J Clin Epidemiol* 2000, **53**:285-289.

28. Jones RN, Gallo JJ: **Education and sex differences in the mini-mental state examination: Effects of differential item functioning.** *Journals of Gerontology Series B-Psychological Sciences & Social Sciences* 2002, **57**: P548-58.
29. Mungas D, Reed BR, Crane PK, Haan MN, Gonzalez H: **Spanish and English Neuropsychological Assessment Scales (SENAS): Further development and psychometric characteristics.** *Psychol Assess* 2004, **16**:347-359.
30. Niti M, Ng TP, Chiam PC, Kua EH: **Item response bias was present in instrumental activity of daily living scale in Asian older adults.** *Journal of Clinical Epidemiology* 2007, **60**:366-374.
31. Jones RN: **Racial bias in the assessment of cognitive functioning of older adults.** *Aging & Mental Health* 2003, **7**:83-102.
32. Kristensen TS, Bjorner JB, Christensen KB, Borg V: **The distinction between work pace and working hours in the measurement of quantitative demands at work.** *Work Stress* 2004, **18**:305-322.
33. Holland PW, Wainer H: **Differential item functioning.** Hillsdale, New Jersey: Lawrence Erlbaum Associates 1993.
34. Benson J, Hutchinson SR: **The state of the art in bias research in the United States.** *European Review of Applied Psychology* 1997, **47**:281-294.
35. Clauser BE, Mazor KM: **Using statistical procedures to identify differentially functioning test items.** *Educational Measurement: Issues and Practice* 1998, **2**:31-44.
36. Groenvold M, Petersen MA: **The role and use of differential item functioning (DIF) analysis of quality of life data from clinical trials.** *Assessing Quality of Life in Clinical Trials* Oxford: Oxford University PressFayers P, Hays R 2005, 195-208.
37. Teresi JA: **Overview of quantitative measurement methods: Equivalence, invariance, and differential item functioning in health applications.** *Med Care* 2006, **44**:S39-S49.
38. Teresi JA: **Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics.** *Med Care* 2006, **44**:S152-S170.
39. Thissen D, Steinberg L, Wainer H: **Detection of differential item functioning using the parameters of item response models.** *Differential Item Functioning* Hillsdale, New Jersey: Lawrence Erlbaum AssociatesHolland PW, Wainer H 1993, 67-114.
40. Teresi JA, Kleinman M, Ocepek-Welikson K: **Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures.** *Stat Med* 2000, **19**:1651-1683.
41. Millsap RE: **Comments on methods for the investigation of measurement bias in the mini-mental state examination.** *Med Care* 2006, **44**:S171-S175.
42. Angoff WH: **Perspectives on Differential Item Functioning.** *Differential Item Functioning* Holland PW, Wainer H 1993, 3-24.
43. Dorans NJ, Holland PW: **DIF detection and description: Mantel-Haenszel and standardization.** *Differential Item Functioning* Hillsdale, New Jersey: Lawrence Erlbaum AssociatesHolland PW, Wainer H 1993, 35-66.
44. Shealy RT, Stout WF: **An item response theory model for test bias and differential item functioning.** *Differential Item Functioning* Hillsdale, New Jersey: Lawrence Erlbaum AssociatesHolland PW, Wainer H 1993, 197-240.
45. Fleishman JA: **Using MIMIC models to assess the influence of differential item functioning.** Presented at the Advances in Health Outcomes Measurement conference, Washington DC 2004 [http://www.outcomes.cancer.gov/conference/irt/fleishman.pdf].
46. Swaminathan H, Rogers HJ: **Detecting differential item functioning using logistic regression procedures.** *Journal of Educational Measurement* 1990, **27**:361-370.
47. Rogers HJ, Swaminathan H: **A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning.** *Applied Psychological Measurement* 1993, **17**:105-116.
48. French AW, Miller TR: **Logistic regression and its use in detecting differential functioning in polytomous items.** *Journal of Educational Measurement* 1996, **33**:315-332.
49. Jodoin MG, Gierl MJ: **Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection.** *Applied Measurement in Education* 2001, **14**:329-349.
50. Scott SC, Goldberg MS, Mayo NE: **Statistical assessment of ordinal outcomes in comparative studies.** *J Clin Epidemiol* 1997, **50**:45-55.
51. Shimizu Y, Zumbo BD: **A logistic regression for differential item functioning primer.** *Japan Language Testing Association Journal* 2005, **7**:110-124.
52. Teresi J: **Differential item functioning and health assessment.** Presented at the Advances in Health Outcomes Measurement conference, Washington DC 2004 [http://www.outcomes.cancer.gov/conference/irt/teresi.pdf].
53. Millsap RE, Everson HT: **Methodology review - statistical approaches for assessing measurement bias.** *Applied Psychological Measurement* 1993, **17**:297-334.
54. Crane PK: **Commentary on comparing translations of the EORTC QLQ-C30 using differential item functioning analyses.** *Quality of life research* 2006, **15**:1117-1118.
55. Lai JS, Teresi J, Gershon R: **Procedures for the analysis of differential item functioning (DIF) for small sample sizes.** *Eval Health Prof* 2005, **28**:283-294.
56. Scott NW, Fayers PM, Aaronson NK, Bottomley A, de Graeff A, Groenvold M, Gundy C, Koller M, Petersen MA, Sprangers MAG: **A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales.** *J Clin Epidemiol* 2009, **62**:288-295.
57. Roussos L, Stout W: **A multidimensionality-based DIF analysis paradigm.** *Applied Psychological Measurement* 1996, **20**:355-371.
58. Lewis C: **A note on the value of including the studied item in the test score when analyzing test items for DIF.** *Differential Item Functioning* Hillsdale, New Jersey: Lawrence Erlbaum AssociatesHolland PW, Wainer H 1993, 317-320.
59. Navas-Ara MJ, Gómez-Benito J: **Effects of ability scale purification on the identification of DIF.** *European Journal of Psychological Assessment* 2002, **18**:9-15.
60. Hidalgo-Montesinos MD, Gómez-Benito J: **Test purification and the evaluation of differential item functioning with multinomial logistic regression.** *European Journal of Psychological Assessment* 2003, **19**:1-11.
61. Stump TE, Monahan P, McHorney CA: **Differential item functioning in the short portable mental status questionnaire.** *Res Aging* 2005, **27**:355-384.
62. Crane PK, van Belle G, Larson EB: **Test bias in a cognitive test: Differential item functioning in the CASI.** *Stat Med* 2004, **23**:241-256.
63. Crane PK, Cetin K, Cook KF, Johnson K, Deyo R, Amtmann D: **Differential item functioning impact in a modified version of the Roland-Morris disability questionnaire.** *Quality of Life Research* 2007, **16**:981-990.
64. Crane PK, Hart DL, Gibbons LE, Cook KF: **A 37-item shoulder functional status item pool had negligible differential functioning.** *J Clin Epidemiol* 2006, **59**:478-484.
65. Scott NW, Fayers PM, Aaronson NK, Bottomley A, de Graeff A, Groenvold M, Gundy C, Koller M, Petersen MA, Sprangers MAG: **Interpretation of differential item functioning (DIF) analyses using external review.** *Expert Reviews in Pharmacoeconomics and Outcomes Research* 2010, **10**:253-258.
66. Bender R, Lange S: **Adjusting for multiple testing - when and how?** *Journal of Clinical Epidemiology* 2001, **54**:343-349.
67. Gierl MJ, Khaliq SN: **Identifying sources of differential item functioning on translated achievement tests: a confirmatory analysis.** Presented at the Annual meeting of the National Council on Measurement in Education, New Orleans 2000.
68. Gierl MJ, Rogers WT, Klinger DA: **Using statistical and judgmental reviews to identify and interpret translation differential item functioning.** *Alberta Journal of Educational Research* 1999, **45**:353-376.
69. Hidalgo MD, Lopez-Pina JA: **Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures.** *Educational and Psychological Measurement* 2004, **64**:903-915.
70. Zieky M: **Practical questions in the use of DIF statistics in test development.** *Differential Item Functioning* Hillsdale, New Jersey: Lawrence Erlbaum AssociatesHolland PW, Wainer H 1993, 337-348.
71. Crane PK, Gibbons LE, Ocepek-Welikson K, Cook K, Cella D, Narasimhalu K, Hays RD, Teresi JA: **A comparison of three sets of criteria for determining the presence of differential item functioning.** *Quality of Life Research* 2007, **16**:S69-S84.
72. Borsboom D: **When does measurement invariance matter?** *Med Care* 2006, **44**:S176-S181.
73. Hambleton RK: **Good practices for identifying differential item functioning.** *Med Care* 2006, **44**:S182-S188.
74. Crane PK, Gibbons LE, Narasimhalu K, Lai J-, Cella D: **Rapid detection of differential item functioning in assessments of health-related quality of life: The functional assessment of cancer therapy.** *Quality of Life Research* 2007, **16**:101-114.
75. Hart DL, Deutscher D, Crane PK, Wang Y: **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* 2009, **18**:1067-1083.
76. McHorney CA, Fleishman JA: **Assessing and understanding measurement equivalence in health outcome measures.** *Med Care* 2006, **44**:S205-S210.
77. Scott NW, Fayers PM, Aaronson NK, Bottomley A, de Graeff A, Groenvold M, Gundy C, Koller M, Petersen MA, Sprangers MAG: **The practical impact of differential item functioning analyses in a health-related quality of life instrument.** *Quality of Life Research* 2009, **18**:1125-1130.
78. Langer MM, Hill CD, Thissen D, Burwinkle TM, Varni JW, DeWalt DA: **Item response theory detected differential item functioning between healthy and ill children in quality-of-life measures.** *Journal of Clinical Epidemiology* 2008, **61**:268-276.
79. Engelhard G, Davis M, Hansche L: **Evaluating the accuracy of judgments obtained from item review committees.** *Applied Measurement in Education* 1999, **12**:199-210.
80. Ryan KE, Bachman LF: **Differential item functioning on two tests of EFL proficiency.** *Language Testing* 1992, **9**:12-29.
81. Allalouf A, Hambleton R, Sireci S: **Identifying the causes of translation DIF on verbal items.** *Journal of Educational Measurement* 1999, **36**:185-198.
82. Ercikan K: **Disentangling sources of differential item functioning in multilanguage assessments.** *International Journal of Testing* 2002, **2**:199-215.
83. Huang CD, Church AT, Katigbak MS: **Identifying cultural differences in items and traits - differential item functioning in the NEO personality inventory.** *Journal of Cross-Cultural Psychology* 1997, **28**:192-218.
84. Sireci SG, Berberoglu G: **Using bilingual respondents to evaluate translated-adapted items.** *Applied Measurement in Education* 2000, **13**:229-248.
85. Schmitt AP, Holland PW, Dorans NJ: **Evaluating hypotheses about differential item functioning.** *Differential Item Functioning* Hillsdale, New Jersey: Lawrence Erlbaum AssociatesHolland PW, Wainer H 1993, 281-316.
86. Ramirez M, Teresi JA, Holmes D, Gurrland B, Lantigua R: **Differential item functioning (DIF) and the mini-mental state examination (MMSE).** *Med Care* 2006, **44**:S95-S106.

doi:10.1186/1477-7525-8-81
Cite this article as: Scott *et al.*: Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression. *Health and Quality of Life Outcomes* 2010 **8**:81.