Quality of Life Research (2006) 15: 1103–1115
DOI 10.1007/s11136-006-0040-x

Comparing translations of the EORTC QLQ-C30 using differential item functioning analyses

N.W. Scott[1], P.M. Fayers[1], A. Bottomley[2], N.K. Aaronson[3], A de Graeff[4], M. Groenvold[5,6], M. Koller[7], M.A. Petersen[5] & M.A.G. Sprangers[8] on behalf of the EORTC and the Quality of Life Cross-Cultural Meta-Analysis Group

[1]*Department of Public Health, University of Aberdeen, Foresterhill, Aberdeen, UK (E-mail: p.fayers@abdn.ac.uk);* [2]*Quality of Life Unit, European Organisation for Research and Treatment of Cancer Data Center, Brussels, Belgium;* [3]*Division of Psychosocial Research and Epidemiology, Netherlands Cancer Institute, Amsterdam, The Netherlands;* [4]*Division of Medical Oncology, Department of Internal Medicine, University Medical Centre, Utrecht, The Netherlands;* [5]*Department of Palliative Medicine, Bispebjerg Hospital, Copenhagen, Denmark;* [6]*Institute of Public Health, University of Copenhagen, Copenhagen, Denmark;* [7]*Centre for Clinical Studies, University Hospital Regensburg, Regensburg, Germany;* [8]*Department of Medical Psychology, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands*

Accepted in revised form 4 January 2006

Abstract

The European Organisation for Research and Treatment of Cancer (EORTC) QLQ-C30 is one of the most widely used quality of life instruments for cancer patients. The aim of this study was to assess whether there were linguistic differences in the way an international sample answered the EORTC QLQ-C30 questionnaire. Thirteen translations of the EORTC QLQ-C30, representing 22 countries, were investigated using a database of 27,891 respondents, incorporating 103 separate studies. Differential item functioning (DIF) analyses were conducted using logistic regression to identify items which, after controlling for subscale, were answered differently by language of administration. Both uniform and non-uniform DIF were assessed. Although most languages showed similar results to English, at least one instance of statistically significant DIF was identified for each translation, and a few of these differences were large. In some cases, the patterns were supported by the results of qualitative interviews with bilingual people. Although, overall, there appeared to be good linguistic equivalence for most of the EORTC QLQ-C30 items, several scales showed strongly discrepant results for some translations. Some of these effects are large enough to impact on the results of clinical trials. Based on our experience in this study, we suggest that validation of translations of health-related quality of life instruments should include exploration of DIF.

Key words: Cancer, Cross-cultural research, Differential item functioning, EORTC QLQ-C30, Translations

Introduction

The European Organisation for Research and Treatment of Cancer (EORTC) QLQ-C30 [1–3] is one of the most widely used instruments designed to assess the quality of life of cancer patients [4]. It is now available in 59 languages and has been registered for use in over 4000 studies around the world.

The core QLQ-C30 questionnaire includes 30 items comprising five functional scales (physical (PF), role (RF), emotional (EF), cognitive (CF) and social (SF) functioning), three symptom scales

(fatigue (FA), nausea/vomiting (NV) and pain (PA)), six single-item symptom items and one scale measuring global health status/quality of life (QL). These scales comprise between two and five items each. Since its general release in 1993, there have been four versions of the EORTC QLQ-C30 and new versions of three scales (QL2, PF2 and RF2) have been created [5]. In the current version of the questionnaire all items have four response categories (not at all, a little, quite a bit and very much), except for the two items of the QL2 scale which use a seven-point scale.

In contrast to instruments that are initially developed in one country or language only, the EORTC Quality of Life Group has from the outset employed a parallel approach and individuals from many different countries have been involved in all stages of the development of the EORTC QLQ-C30 questionnaire [6]. Although the EORTC QLQ-C30 has been used in numerous international studies it is not certain to what extent translation and cultural differences influence response. Each translation of the questionnaire is produced using a rigorous process of forward and back translations followed by extensive pilot testing [7], but even the best possible translation may not be exactly equivalent to the original English version. In addition, even if the translation is perfect, cultural differences between countries could result in markedly different response patterns, even after controlling for quality of life status. Such effects have the potential to influence the results of international studies such as multicentre clinical trials, where trial data from several countries are frequently pooled together.

Differential item functioning (DIF) analyses originated in educational testing settings where they are commonly used to determine whether individual test items are unfair to particular groups (e.g. females or ethnic minorities) even after allowing for overall test ability [8]. Many different approaches for the conduct of DIF analyses are available including contingency table, logistic regression and item response theory techniques [9–12]. DIF methodology is now starting to be used in other fields, including for testing comparability of translations [13–17], although DIF analyses of quality of life instruments are still relatively uncommon.

An earlier report by the EORTC Quality of Life Group found statistically significant DIF for some translations of the EF scale of the EORTC QLQ-C30 [18]. The following paper aims to use DIF analyses to identify differences in international response to all multi-item scales.

Methods

Datasets

One hundred and three datasets were received from a wide variety of sources. Anonymised data from 52 international clinical trials and other studies using the EORTC QLQ-C30 were received from the EORTC Data Center with permission from the chairs of the relevant EORTC clinical groups and with the approval of the EORTC Board. A further 51 datasets were received from individuals and organisations from around the world. Although most datasets were studies of cancer patients, four of the largest (from Norway, Denmark, Germany and Austria) were surveys of the general population. The following information was extracted for all respondents in each study when available: country, translation used, primary disease site, cancer stage, age, gender, ethnic group and performance status. Only baseline (pre-treatment) data were included, but data were also used if treatment status was not known. In the majority of cases, the translation of the EORTC QLQ-C30 had to be inferred from the geographical location of the hospital involved. Data from all four versions of the EORTC QLQ-C30 were included when possible, although only results for the current versions of the subscales are reported.

The analyses in this paper are restricted to data for 27,891 respondents representing the 13 translations with available baseline data for at least 200 people. The original English version was used in the United Kingdom, the United States, Canada, Australia, Singapore, Myanmar, New Zealand and Ireland; the Italian translation in Italy and Switzerland; the French translation in France, Belgium, Canada and Switzerland; the Dutch translation in the Netherlands and Belgium; and

the German translation in Germany, Austria and Switzerland. The French Canadian translation differs from the standard French version for some items; data from this version were therefore only used for the QL2, EF, FA, NV and PA scales. Two different Chinese translations are represented, one used in Singapore and one in Taiwan. A further six translations of the EORTC QLQ-C30 were represented by a single country (Table 1).

Analyses

DIF analyses were conducted for the nine multi-item subscales of the EORTC QLQ-C30 using Stata version 9.0. For consistency of interpretation the coding of items 29 and 30, the only positively phrased questions in the EORTC QLQ-C30, was reversed so that higher scores represented worse quality of life for these items also. For every item within each subscale a single ordinal logistic regression model predicting the item response (using the proportional odds model [19]) was used to derive a common odds ratios for each translation. Variables representing each translation were entered into the model simultaneously with English used as the reference category. The overall scale score was included in the model as the "matching" variable and analyses were also adjusted for age, gender, cancer site and stage of disease (using the four categories: no cancer, Stage I–III, Stage IV/recurrent/metastatic and not known). It was not possible to adjust analyses for performance status or ethnic group as this information was often unavailable. Cancer site was classified using 18 categories and respondents with unknown site were included in an "other site/not known" category. Respondents with missing age or gender were excluded from the analyses. Binary logistic regression (using "not at all" versus the other three categories) was used instead of ordinal regression for the PF2 scale because for some items relatively few respondents chose the "quite a bit" and "very much" categories. Danish was excluded from the QL2, PF2 and RF2 analyses because most of the available data for this translation used the earlier versions of these scales.

Table 1. Amount of available data by country

Translation	Country	Number of respondents
Chinese (Singapore)	Singapore	258
Chinese (Taiwan)	Taiwan	208
Danish	Denmark	795
Dutch	Netherlands	3253
	Belgium	559
French	France	1426
	Belgium	117
	Canada	111
	Switzerland	71
German	Germany	7056
	Austria	1450
	Switzerland	200
Italian	Italy	656
	Switzerland	19
Norwegian	Norway	4440
Polish	Poland	373
Spanish	Spain	917
Swedish	Sweden	994
Turkish	Turkey	242
English	United Kingdom	2018
	United States	1326
	Canada	537
	Australia	401
	Singapore	262
	Myanmar	103
	New Zealand	72
	Ireland	27
Total		27,891

For selected items graphs show log odds ratios for each translation derived from the logistic regression analyses with their 95% confidence intervals. Log odds ratios greater than zero mean that respondents using that translation were more likely to assign a higher score (indicating worse health outcome) to that item compared with English and relative to other items in the same scale. Log odds ratios less than zero imply that respondents using that language were less likely to score highly on this item.

The above analyses are only designed to detect uniform DIF; non-uniform DIF occurs when the direction and magnitude of DIF effects vary according to the level of the overall scale score. In this study the presence of non-uniform DIF was assessed by adding interaction terms defined as the product of each language variable and the overall scale score into each logistic regression model.

Interpretation

Because of the large number of statistical tests conducted, a stricter cut-off for statistical

significance was used to confirm the presence of both uniform and non-uniform DIF ($p < 0.001$). When interpreting the results, however, it is not sufficient to consider whether the results are statistically significant; it should also be determined whether the magnitude of any DIF effect is of practical importance. Various systems exist for assessing the size of uniform DIF effects [13, 20, 21]; we have used the convention that in order for a result to be considered statistically significant the absolute value of the log odds ratio should be greater than 0.64 as well as having a p-value less than 0.001 [13, 18]. Vertical lines have therefore been marked on the graphs at 0.64 and −0.64. The use of this cut-point, however, is somewhat arbitrary and results are dependent on which language is used as the reference category. In some situations, it may be the reference category, English, that is inconsistent with other translations. To assess the magnitude of non-uniform DIF the difference in pseudo-R^2 between two separate models was considered for each combination of item and translation: one including just the overall scale score and translation and one also including the interaction between these two variables. Non-uniform DIF effects were only considered significant if (a) the p-value was less than 0.001, and (b) the difference in R^2 was at least 0.035 [22].

Because each subscale contains no more than five items it is important to realise that some DIF effects may be due to "pseudo-DIF", e.g. if there was a translation problem for only one item of a two-item scale the results will typically show DIF for both items in opposite directions and it will not be possible to tell which item (or items) is causing the DIF. Therefore, DIF results should not be considered in isolation and the results for all the items that make up a scale should be interpreted together.

In addition, it is not possible to tell what is the cause of a significant DIF effect from the statistical results alone. To help explore whether DIF effects were caused by translation issues we conducted a number of structured interviews with bilingual people. For each item of the questionnaire, interviewees were asked to state whether a hypothetical group of bilingual people would tend to obtain higher or lower scores when using the translated version compared with the English version. The equivalence of the translation of each item was assessed using a seven-point Likert scale and qualitative comments were also recorded. As the number of interviewees per language was small, this was regarded as an exploratory exercise only.

Results

Table 1 shows the amount of available data by country for each translation of the questionnaire. Table 2 gives details of the gender, cancer site, stage of disease and age distribution of this sample.

Interviews with bilingual people

Structured interviews were conducted with 40 bilingual people. Between one and six (median of four) people were interviewed for each translation

Table 2. Characteristics of included respondents (n = 27,891)

Age		
Mean: 57.0 (SD 14.7)		Not known: 703 (2.5%)
Gender		
Male	15,616	(58.7%)
Female	11,009	(41.3%)
Not known	1266	
Stage		
No cancer	8170	(31.9%)
I–III	9326	(36.4%)
IV/metastatic/recurrent	8128	(31.7%)
Not known	2267	
Cancer site		
No cancer	7804	(28.6%)
Prostate	3232	(11.8%)
Lung	2974	(10.9%)
Head and neck	2123	(7.8%)
Breast	2092	(7.7%)
Oesophagus/stomach	1744	(6.4%)
Colorectal	1667	(6.1%)
Malignant melanoma	1058	(3.9%)
Gynaecological	1036	(3.8%)
Myeloma	938	(3.4%)
Liver/bile/pancreas	596	(2.2%)
Other cancer	573	(2.1%)
Malignant lymphoma	394	(1.4%)
Testicular	349	(1.3%)
Brain	259	(0.9%)
Genito-urinary	227	(0.8%)
Leukaemia	204	(0.7%)
Sarcoma	54	(0.2%)
Not known	567	

considered in this report. The majority of interviewees were staff or students at the University of Aberdeen, UK.

Results of the DIF analyses

DIF results for each subscale of the EORTC QLQ-C30 are presented separately. In general only translations meeting both significance criteria for uniform DIF (p-value less than 0.001 and a log odds ratio coefficient with absolute value greater than 0.64) will be commented on in the text. A summary of the DIF results is provided in Table 3. Possible interpretation of the results derived from the interviews is discussed in the text but only if the same comment was made by at least two interviewees. Turkish translations could not be flagged in this way because only one interview was available.

Global Health Status/Quality of Life (revised version) (QL2) (two items)

For most languages the log odds ratios were similar, suggesting that there were few international differences in how these two questions were answered. Although Italian speakers tended to have relatively worse overall health and better quality of life compared with English speakers, no translations met the criteria for significant DIF (Figure 1).

Table 3. Summary of uniform and non-uniform DIF results for each item by translation

Scale	Item	ZH(Sin)	ZH(Tai)	DA	NL	FR	DE	IT	NO	PL	ES	SV	TR
QL2	Q29			nc									
	Q30			nc									
PF2	Q1			nc								−	
	Q2			nc						−			
	Q3			nc	+				+				+
	Q4		+	nc	−			+	+	+	−	+	+
	Q5			nc									+
RF2	Q6			nc						+			
	Q7			nc			−						
EF	Q21	−								−			
	Q22	−	−				+		−				−
	Q23				−						−		
	Q24	+	+						+	+		+	
CF	Q20				+							+	
	Q25				−							−	
SF	Q26			+			+						
	Q27			−			−				−		
FA	Q10										−		
	Q12							+			+		+
	Q18		−						−				−
NV	Q14	−						−					
	Q15							+					
PA	Q9					−	−		−				
	Q19		+			+	+						

Translations: Singapore Chinese (ZH(Sin)), Taiwan Chinese (ZH(Tai)), Danish (DA), Dutch (NL), French (FR), German (DE), Italian (IT), Norwegian (NO), Polish (PL), Spanish (ES), Swedish (SV), Turkish (TR).
"+" indicates that speakers of that language were more likely to report symptoms for that item compared with English and with other items in the same scale ($p < 0.001$ and $|\text{log odds ratio}| > 0.64$). "−" indicates that speakers of that language were less likely to score highly on that item. "nc" indicates that DIF analyses were not conducted because of insufficient sample size.
Shaded cells indicate items with statistically significant ($p < 0.001$) non-uniform DIF. None of these items also met the magnitude criterion (R^2 change of at least 0.035).

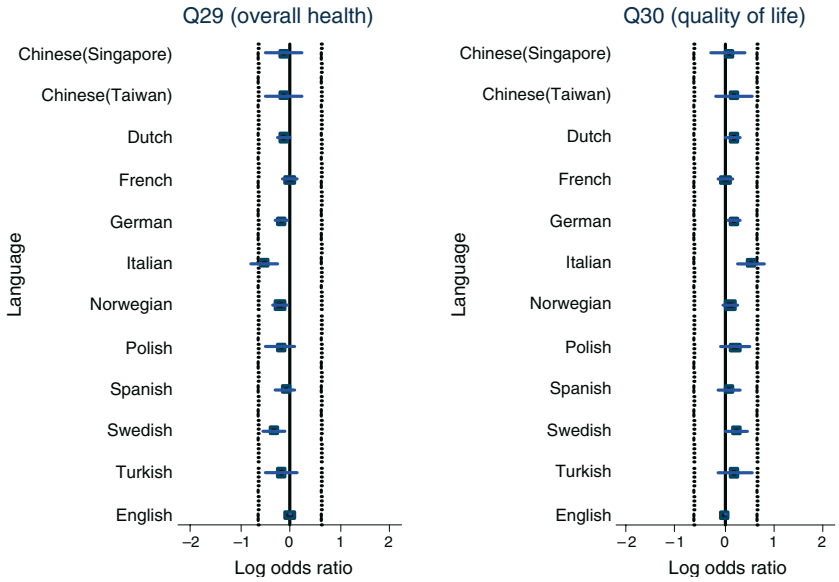

Figure 1. Global Health Status/Quality of Life (version 2) (QL2).

Physical Functioning (revised version) (PF2) (five items)

For this scale there were a number of instances of translations with statistically significant DIF which also met the criterion of having absolute log odds ratios above 0.64 (Figure 2). Swedish respondents were less likely to score highly on Q1 (Do you have any trouble doing strenuous activities?) compared with English speakers, whereas Polish speakers tended to have relatively fewer problems with taking a long walk (Q2). Dutch, Norwegian and Turkish speakers reported relatively more difficulties compared with English speakers for the question about taking a short walk (Q3). Q4 (Do you need to stay in bed or a chair during the day?) showed the most variation of all the questions in this scale. Dutch and Spanish speaking respondents were less likely to score highly; those using the Taiwan Chinese, Italian, Norwegian, Polish, Swedish and Turkish translations were relatively more likely to report needing to stay in bed or a chair. Finally, Turkish speakers were relatively more likely to report needing help with eating, dressing, washing or using the toilet (Q5).

All three Swedish interviewees thought that Swedes would report higher scores for Q4 because the Swedish translation of this item was less likely to imply having to sit or lie down due to ill health.

Role Functioning (revised version) (RF2) (two items)

There was some evidence that respondents using German or Polish tended to have relatively more limitations with work (Q6) and fewer limitations with hobbies (Q7) than English speakers (data not shown).

Emotional Functioning (EF) (four items)

A number of instances of significant DIF were observed compared with English (Figure 3). Respondents using the Polish or Singapore Chinese translations tended to score lower on Q21 (Did you feel tense?). Respondents using the Norwegian, Turkish or either of the Chinese translations tended to score lower on Q22 (Did you worry?) and those using German tended to report worrying more often. Those using Dutch or Spanish tended to score lower on Q23 (Did you feel irritable?). Finally, respondents using the Norwegian, Swedish, Polish or either of the Chinese versions tended to score relatively higher on Q24 (Did you feel depressed?).

Three possible linguistic interpretations of the DIF results were identified from the interviews

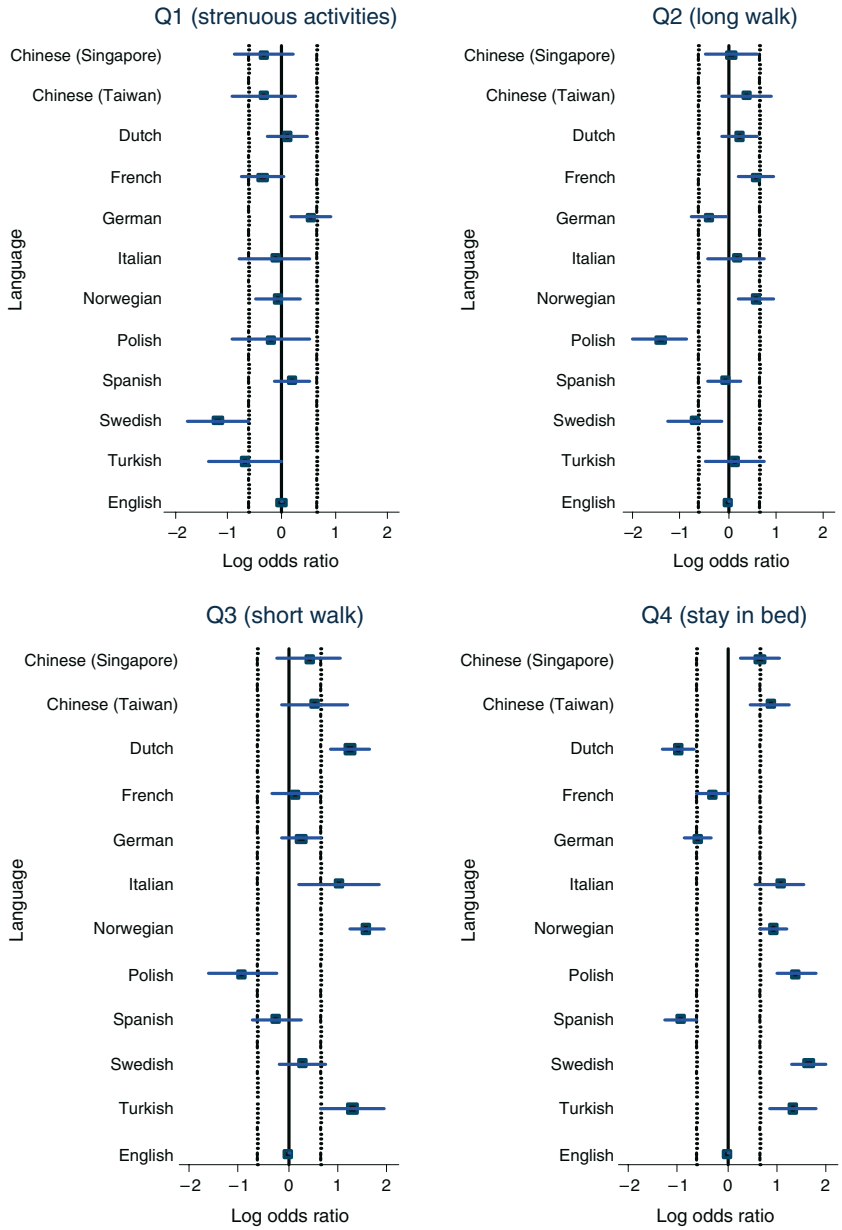

Figure 2. Physical Functioning (version 2) (PF2). (Continued on p. 1110.)

with bilingual people, two of which have been previously identified [18]. Two out of the three Norwegian interviewees mentioned that the Norwegian translation of Q22 may be a stronger statement than English and may have a meaning closer to "anxious" than "worry". All three Swedish interviewees reported that the word for "depressed" in the Swedish translation of Q24 was weaker than English and could mean just "feeling down". Similarly, two out of four Polish interviewees judged the Polish translation of Q24 to be a weaker statement than English.

Cognitive Functioning (CF) (two items)

Respondents using the Dutch and Swedish translations tended to report relatively more difficulties with concentrating (Q20) than with remembering

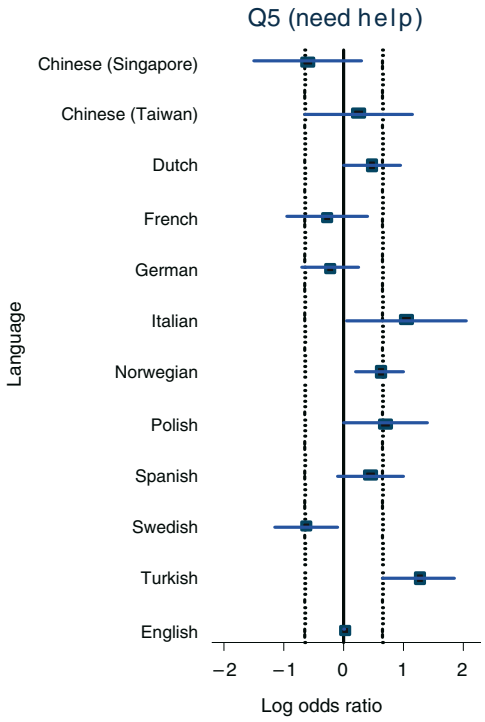

Figure 2. Continued.

things (Q25) compared with English speakers (data not shown).

Social Functioning (SF) (two items)

Respondents using the Danish, German and Spanish translations tended to score relatively higher on the question about family life (Q26) and lower on the question about social activities (Q27) (Figure 4).

Fatigue (FA) (three items)

Spanish-speaking respondents tended to score relatively lower on Q10 (Did you need to rest?) compared with English speakers (Figure 5). Those using the Italian, Spanish and Turkish translations were relatively more likely to report feeling weak (Q12). Those using the Norwegian, Turkish and Taiwan Chinese versions tended to score lower on Q18 (Were you tired?).

Two bilingual interviewees thought that the lower scores for Spanish speakers on Q10 may be due to the translation used: it means literally "Did you stop to rest?".

Nausea/vomiting (NV) (two items)

There was evidence that respondents using the Italian and Singapore Chinese translations were less likely to score highly on Q14 (Have you felt nauseated?) compared with English and relative to Q15 (Have you vomited?) (data not shown).

Pain (PA) (two items)

For this scale the reference language (English) was the most extreme with relatively lower scores on Q19 (Did pain interfere with your daily activities?) compared with Q9 (Have you had pain?). There was some evidence that, compared with English speakers, those using the French, German, Norwegian and Taiwan Chinese translations tended to score relatively lower on Q9 and higher on Q19 (data not shown).

Non-uniform DIF

Twenty-five instances of statistically significant non-uniform DIF (using $p < 0.001$) were observed, nine of which involved the German translation (Table 3). None of the 25 instances of non-uniform DIF, however, met the magnitude criterion since all had an R^2 difference of less than 0.035.

Discussion

The results show that although the results for most countries tended to be similar there was some evidence for DIF in at least one scale for each of the translations examined.

A variety of different methods have been used to detect DIF in the literature. We chose to use logistic regression modelling due to its flexibility and because this method has been shown to perform similarly to other methods for the detection of uniform and non-uniform DIF [18, 23–25]. A single model could be used for each item and analyses could be adjusted for a number of other possible confounding variables.

There is still considerable debate about how to interpret the results of DIF analyses. There is a strong consensus that items should only be flagged as having DIF if they meet magnitude as well as statistical significance criteria, but different criteria

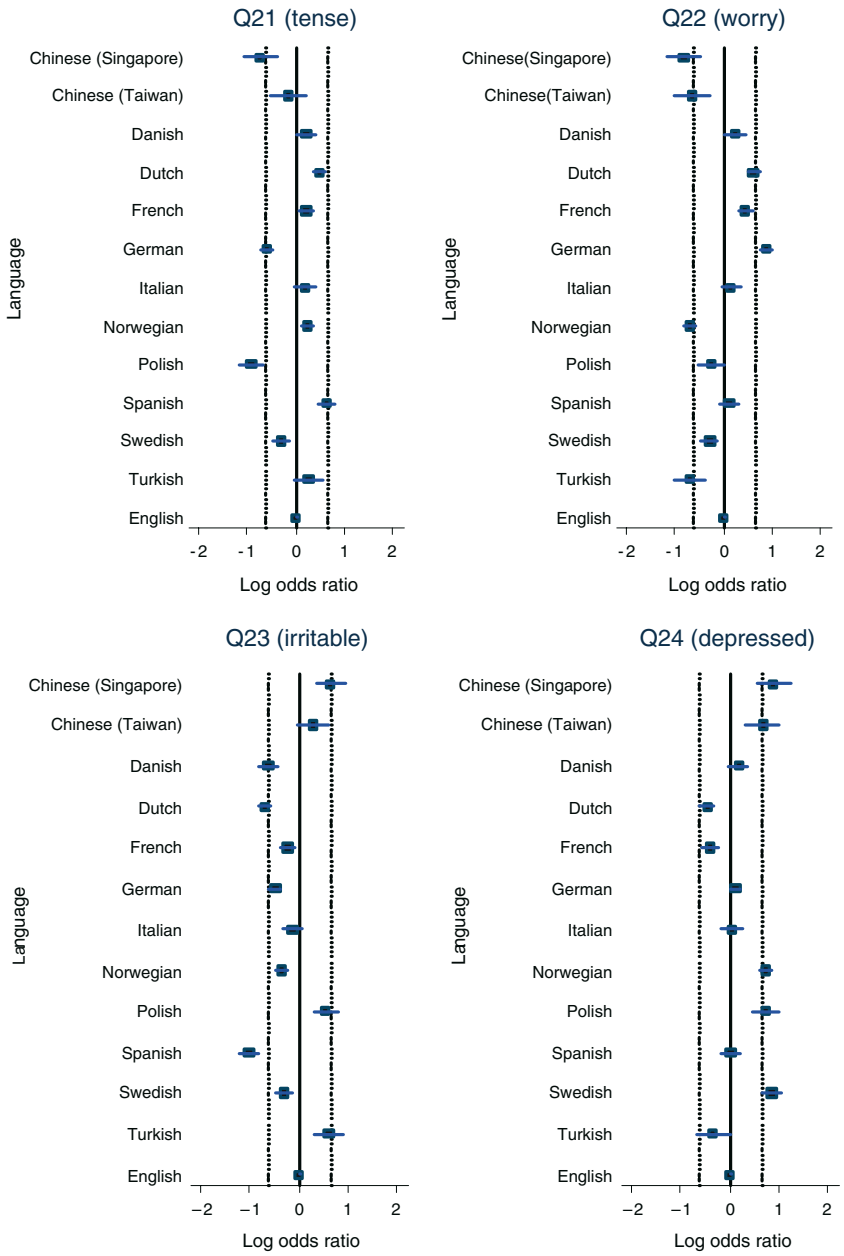

Figure 3. Emotional Functioning (EF).

may result in very different numbers of items being identified [26]. Given the large number of hypothesis tests involved in our study we chose to use a very conservative significance level of 0.001 and to comment only on translations attaining this level and meeting a magnitude criterion of having absolute log odds ratios greater than 0.64. A similar double significance criterion was applied for non-uniform DIF. Although there were no instances of non-uniform DIF meeting the magnitude criterion, the cut-offs are currently still a matter of debate; as well as the 0.035 value used in this paper, an R^2 difference cut-off of 0.13 has also

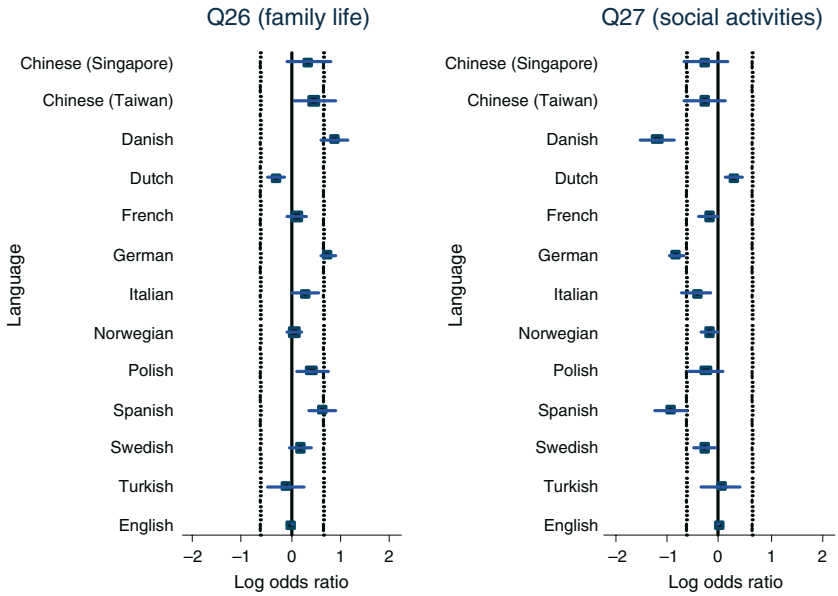

Figure 4. Social Functioning (SF).

been proposed [20] but in our study no items would have met either of these criteria.

If significant uniform DIF is present for an item then the interpretation of this result is not straightforward. It could be due to a number of reasons including influence of confounding factors, pseudo-DIF caused by another item in that scale, translation inequivalence (either for the wording of the item or for the response category labels) or it could be because of genuine cultural differences in response. In addition, the studies that contributed to the analyses were diverse and although we adjusted analyses for age, gender, cancer site and stage of disease the DIF results may reflect other characteristics of those included. Data for some of the languages in this report came from a relatively small number of studies and it is possible that some of our results may partly reflect shared characteristics of these patients.

An important additional difference between our study and other DIF studies is that the scales are short and contain between two and five items each. Therefore pseudo-DIF is much more of a problem than for the longer scales more common in other applications such as educational testing. The use of item response theory scoring in logistic regression DIF analyses has been suggested to avoid this issue [12] but we chose to use unweighted scale scoring as this is the standard method used for scoring the EORTC QLQ-C30 [27].

Others have tried to identify reasons for translation DIF using a combination of statistical and judgmental methods [28–31] but the use of substantive analyses to try to identify the underlying causes of statistically identified DIF items has often been unsuccessful [32]. Our exploratory interviews with bilingual people were occasionally useful and suggested some possible translation reasons for the DIF results. Often, however, there was no clear interpretation and sometimes the interviews did not yield results consistent with the DIF analyses.

It can be difficult to quantify how much of a problem the issue of DIF is for users of the QLQ-C30 as translating the log odds ratio into a clinically meaningful scale can only be done in the context of a specific scenario. As an example, we applied our results to a real Norwegian study [33] and calculated the effect on the FA scale score of using the English translation instead of Norwegian, assuming true DIF existed for Q18 only and not for the other two items of this scale. Our findings suggested that this would result in FA scale scores that were around six points higher, which

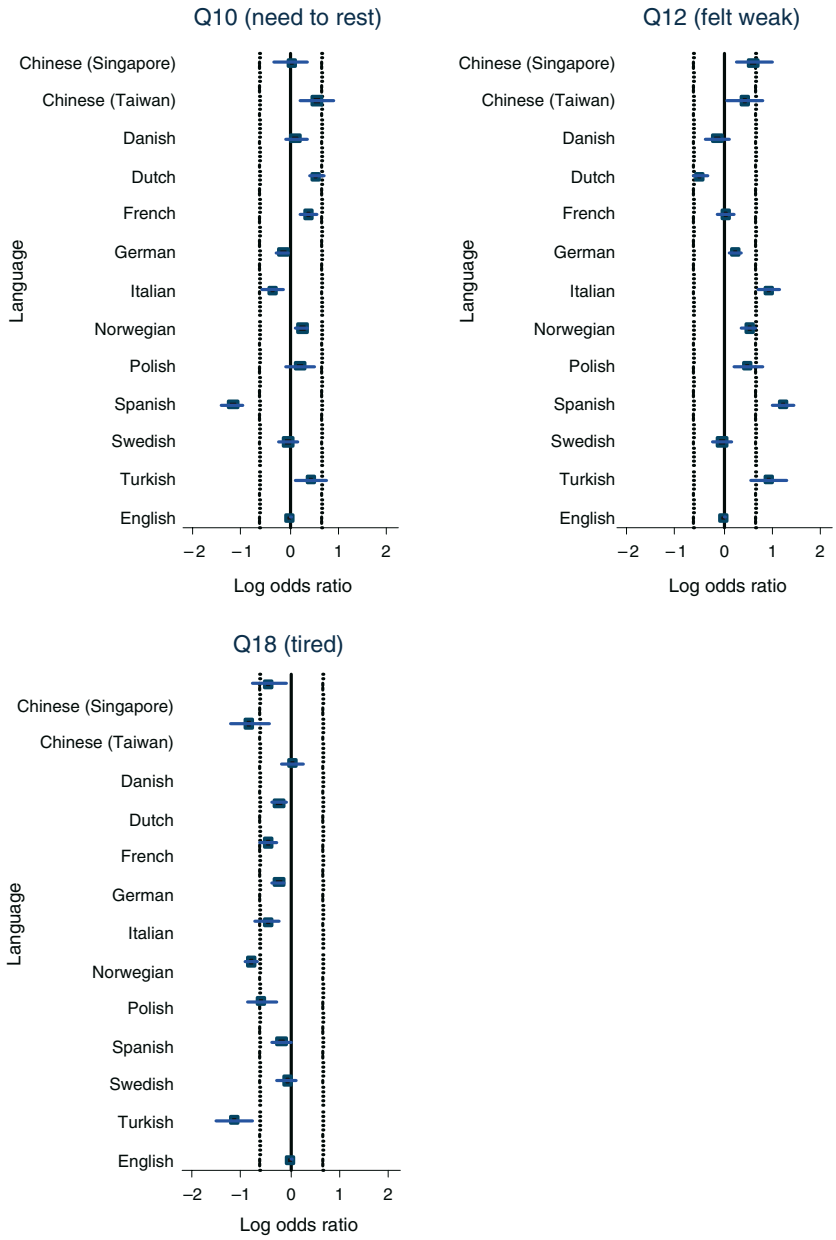

Figure 5. Fatigue (FA).

corresponds to a small but clinically important difference [5]. Although DIF has the potential to bias international comparative studies using the EORTC QLQ-C30, it may be less of a problem in clinical trials, however, as randomised groups are usually stratified by country or centre and biases ought to occur in each group equally.

The issue of DIF may affect all translated instruments but its true extent often remains unidentified. We have successfully conducted extensive DIF analyses using a unique large international database and for the first time we have provided a reasonable picture of the location, direction and number of DIF problems between

translations of the EORTC QLQ-C30. In general the translations were very good, as might be expected because of the rigour of the translation process, but many instances of DIF were still identified. A small number of these effects were large and were substantiated by qualitative interviews. As this might impact on trials and other studies, we plan to review these translations. We also recommend that other groups apply similar methods to evaluate the translations of their quality of life instruments.

Members of the Quality of Life Cross-Cultural Meta-Analysis Group

Australia: M King, S Leutenegger, N Spry; *Austria*: E Greimel, B Holzner; *Belgium*: A Bottomley, C Coens, G de Castro, K West; *Brazil*: C de Souza; *Canada*: A Bezjak, M Whitehead; *Denmark*: M Groenvold, M Klee, M Petersen; *France*: A Brédart, T Conroy, C Rodary; *Germany*: M Koller, O Krauß, T Küchler, B Malchow, R Schwarz; *Greece*: K Mystakidou; *Iran*: A Montazeri; *Italy*: C Brunelli, M Tamburini; *Japan*: H Zhao; *Netherlands*: N Aaronson, A de Graeff, R de Leeuw, M Muller, M Sprangers; *Norway*: K Bjordal, E Brenne, M Hjermstad, M Jordhøy, P Klepstad, S Sundstrøm, F Wisløff; *Singapore*: YB Cheung, SB Tan, J Thumboo, HB Wong; *South Korea*: YH Yun; *Spain* J Arraras; *Sweden*: M Ahlner-Elmqvist; *Switzerland* P Ballabeni, J Bernhard; *Taiwan* W-C Chie; *Turkey* U Abacioglu; *UK* J Blazeby, J Bruce, A Davies, P Fayers, L Friend, Z Krukowski, T Massett, T Matsuoka, J Nicklin, J Ramage, N Scott, A Smyth-Cull, T Young; *USA* D Cella, D-L Esseltine, C Gotay, I Pagano.

Contributing groups

European Organisation for Research and Treatment of Cancer (EORTC) Brain Group, EORTC Breast Cancer Group, EORTC Chronotherapy Group, EORTC Gastro-Intestinal Group, EORTC Genito-Urinary Group, EORTC Gynaecological Group, EORTC Head and Neck Group, EORTC Leukaemia Group, EORTC Lung Group, EORTC Lymphoma Group, EORTC Melanoma Group, EORTC Quality of Life Group, EORTC Radiotherapy Group, EORTC Soft Tissue Group, National Cancer Institute Grant CA60068, National Cancer Institute of Canada (NCIC) Clinical Trials Group, Swiss Group for Clinical Cancer Research (SAKK).

Acknowledgements

We gratefully acknowledge the assistance of the many individuals who helped supply datasets for this study. This work was funded by the EORTC Quality of Life Group and the University of Aberdeen and carried out under the auspices of the EORTC Quality of Life Group.

References

1. Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993; 85(5): 365–376.
2. Aaronson NK, Cull AM, Kaasa S, Sprangers MAG. The European Organization for Research and Treatment of Cancer EORTC) modular approach to quality of life assessment in oncology: an update. In: Spilker B (ed.), Quality of Life and Pharmacoeconomics in Clinical Trials, 2nd ed. Lippincott-Raven, Philadelphia, 1996: 179–189.
3. Fayers P, Bottomley A. Quality of life research within the EORTC – the EORTC QLQ-C30. Eur J Cancer 2002; 38(Suppl 4): S125–S133.
4. Garratt A, Schmidt L, Mackintosh A, Fitzpatrick R. Quality of life measurement: bibliographic study of patient assessed health outcome measures. BMJ 2002; 324(7351): 1417–1419.
5. Fayers P, Aaronson N, Bjordal K, Groenvold M, Curran D, Bottomley A. EORTC QLQ-C30 Scoring Manual. 3rd ed., Brussels: European Organization for Research and Treatment of Cancer; 2001.
6. Aaronson NK. Assessing the quality of life of patients with cancer: East meets West. Eur J Cancer 1998; 34(6): 767–769.
7. Cull A, Sprangers M, Bjordal K, Aaronson N, West K, Bottomley A. EORTC Quality of Life Group translation procedure. Brussels: European Organization for Research and Treatment of Cancer, 2002.
8. Differential item functioning. Holland PW and Wainer H, (eds), Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1993.
9. Clauser BE, Mazor KM. Using statistical procedures to identify differentially functioning test items. Educ Measure Issues Pract 1998; 2: 31–44.
10. Benson J, Hutchinson SR. The state of the art in bias research in the United States. Euro Rev Appl Psychol 1997; 47(4): 281–294.
11. Teresi JA. Statistical methods for examination of differential item functioning (DIF) with applications to

cross-cultural measurement of functional, physical and mental health. J Mental Health Aging 2001; 7(1): 31–40.
12. Crane PK, van Belle G, Larson EB. Test bias in a cognitive test: differential item functioning in the CASI. Stat Med 2004; 23(2): 241–256.
13. 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(11): 1189–1202.
14. 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(3): 355–365.
15. Orlando M, Marshall GN. Differential item functioning in a Spanish translation of the PTSD checklist: detection and evaluation of impact. Psychol Assess 2002; 14(1): 50–59.
16. 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(12): 1271–1278.
17. 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(2): 233–259.
18. Petersen MA, Groenvold M, Bjorner JB, et al. Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire. Qual Life Res 2003; 12(4): 373–385.
19. Scott SC, Goldberg MS, Mayo NE. Statistical assessment of ordinal outcomes in comparative studies. J Clin Epidemiol 1997; 50(1): 45–55.
20. 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. Ottawa: Directorate of Human Resources Research and Evaluation, Department of National Defense, 1999.
21. 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(3): 285–289.
22. Gierl M, Khaliq SN, Boughton K. Gender differential item functioning in mathematics and science: prevalence and policy implications. Presented at the Annual Meeting of the Canadian Society for the Study of Education, Sherbrooke, Quebec, 1999.
23. Rogers HJ, Swaminathan H. A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. Appl Psychol Measure 1993; 17(2): 105–116.
24. Welkenhuysen-Gybels J, Billiet J. A comparison of techniques for detecting cross-cultural inequivalence at the item level. Qual Quant 2002; 36: 197–218.
25. Hidalgo MD, Lopez-Pina JA. Differential item functioning detection and effect size: a comparison obetween logistic regression and Mantel-Haenszel procedures. Educ Psychol Measure 2004; 64(6): 903–915.
26. Jodoin MG, Gierl MJ. Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. Appl Measure Educ 2001; 14(4): 329–349.
27. Petersen MA, Groenvold M, Aaronson N, Brenne E, et al. Scoring based on item response theory did not alter the measurement ability of EORTC QLQ-C30 scales. J Clin Epi 2005; 58: 902–908.
28. Allalouf A, Hambleton R, Sireci S. Identifying the causes of translation DIF on verbal items. J Educ Measure 1999; 36: 185–198.
29. Gierl MJ, Rogers WT, Klinger DA. Using statistical and judgmental reviews to identify and interpret translation differential item functioning. Alberta J Educ Res 1999; 45(4): 353–376.
30. 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.
31. Puhan G, Gierl MJ. Evaluating the comparability of English- and French-speaking examinees on a science achievement test administered using two-stage testing. Presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, 2003.
32. Roussos L, Stout W. A multidimensionality-based DIF analysis paradigm. Appl Psychol Measure 1996; 20: 355–371.
33. Wisloff F, Hjorth M, Kaasa S, Westin J. Effect of interferon on the health-related quality of life of multiple myeloma patients: results of a Nordic randomized trial comparing melphalan-prednisone to melphalan-prednisone $+\alpha$-interferon. Brit J Haematol 1996; 94: 324–332.

Address for correspondence: Peter Fayers, Department of Public Health, University of Aberdeen, Polwarth Building, Foresterhill, Aberdeen, AB25 2ZD, United Kingdom
Phone: +44-1224-559573; Fax: +44-1224-550925
E-mail: p.fayers@abdn.ac.uk