Int. J. Behav. Med. (2009) 16:136–147
DOI 10.1007/s12529-009-9048-2

Cross-Language Differential Item Functioning of the Job Content Questionnaire Among European Countries: The JACE Study

BongKyoo Choi · Jakob Blue Bjorner · Per-Olof Ostergren · Els Clays · Irene Houtman · Laura Punnett · Annika Rosengren · Dirk De Bacquer · Marco Ferrario · Maaike Bilau · Robert Karasek

Published online: 3 July 2009

Abstract
Background Little is known about cross-language measurement equivalence of the job content questionnaire (JCQ)
Purpose The purposes of this study were to assess the extent of cross-language differential item functioning (DIF) of the 27 JCQ items in six languages (French, Dutch, Belgian-French, Belgian-Dutch (Flemish), Italian, and Swedish) from six European research centers and to test whether its effects on the scale-level mean comparisons among the centers were substantial or not.
Method A partial gamma coefficient method was used for statistical DIF analyses where the Flemish JCQ was the reference for other language versions. Additionally, equivalence between the Flemish and Dutch translations was subjected to a judgmental review.
Results On average, 36% to 39% of the total tested items appeared to be cross-language DIF items in the statistical analyses. The judgmental review indicated that half of the DIF items may be associated with translation difference. The impacts of the DIF items on the mean comparisons of the JCQ scales between the centers were non-trivial: underestimated skill discretion (Milan), underestimated decision authority (Leiden), underestimated psychological demands (Milan women), and incomparable coworker support (Gothenburg 95).
Conclusion Cross-language DIF of the JCQ among European countries should be considered in international

B. Choi (✉)
Center for Occupational and Environmental Health,
University of California Irvine,
5201 California Avenue, Suite 100,
Irvine, CA 92617, USA
e-mail: b.choi@uci.edu

J. B. Bjorner
National Research Centre for the Working Environment,
Copenhagen, Denmark

P.-O. Ostergren
Division of Social Medicine and Global Health,
Department of Clinical Sciences, Lund University,
Malmo University Hospital,
Malmo, Sweden

E. Clays · D. De Bacquer · M. Bilau
Department of Public Health, Ghent University,
Ghent, Belgium

I. Houtman
TNO Work & Employment, AS Hoofddorp,
Hoofddorp, The Netherlands

L. Punnett · R. Karasek
Department of Work Environment,
University of Massachusetts Lowell,
Lowell, USA

A. Rosengren
Department of Medicine, Sahlgrenska University Hospital/Ostra,
Gothenburg, Sweden

M. Ferrario
Work and Preventive Medicine, Department of Clinical and
Biological Sciences, University of Insubria at Varese,
Varese, Italy

R. Karasek
Department of Psychology, Copenhagen University,
Copenhagen, Denmark

comparative studies on psychosocial job hazards using JCQ scales.

Keywords Gamma coefficient · Measurement · Translation · Equivalence · Psychosocial

Introduction

The job content questionnaire (JCQ), developed originally in the USA [1], has been one of the most utilized instruments to measure psychosocial job characteristics due to its simplicity, reliability, and validity [2, 3]. As of 2008, it has been translated into twenty-three languages (http://www.jcqcenter.org). Considering the present paucity of adequate global surveillance data on work stress risk factors ([4], p. 5; [5], p. 73), existing large JCQ datasets from many countries can be good information sources for assessing psychosocial job hazards in a global economy. However, the comparability of these datasets should be assured in order to interpret them appropriately.

There are a number of methodological issues or requirements for a cross-cultural study using questionnaires [6–11]. One central issue is measurement equivalence across cultures: (a) "whether research instruments elicit the same conceptual frame of reference in culturally diverse groups" and (b) "whether respondents calibrate the intervals anchoring the measurement continuum in the same manner" ([12], p. 644). Measurement non-equivalence between cultures can be a serious threat to the validity of quantitative cross-cultural comparison studies because it becomes hard to tell whether observed mean differences or similarities are reflecting reality or simply measurement artifacts.

In this paper, measurement non-equivalence was evaluated through tests of differential item functioning (DIF) [13–17]. An item shows DIF if "all respondents at a given level of the attribute measured (at a given index score) do not have equal probability of scoring positively on the item regardless of subgroup membership" ([17], p. 264).

In general, employing a standardized translation procedure such as translation and back-translation with bilingual translators is effective to reduce conceptual difference of an item between cultures [18] but cannot rule out its possibility completely due to relative insensitiveness of back-translation procedure to quality of translation [7, 19, 20]. In many cases, translation is not perfect due to culture-bound wordings of item stem (content) and options (response category) [9, 12, 20–22]. In addition, ambiguous wording of the original item can be amplified through translation in target cultures. All of these can lead to measurement non-equivalence between an original and target cultures through eliciting subtly or substantially differential conceptual frames of reference from respondents [15, 23, 24].

Despite the worldwide use of the JCQ, there have been no international translation validation studies [25]. Few studies have examined cross-language or cross-national DIF of the JCQ statistically and/or qualitatively. Two previous studies [26, 27] suggest that JCQ items as other measures may function differently across countries (languages). Karasek et al. [26] raised a doubt about the consistency of the meanings of the JCQ "psychological demands" items across the USA, Canada, the Netherlands, and Japan. Choi et al. [27] reported that 16 of 22 tested JCQ items (of skill discretion, decision authority, psychological demands, supervisor support, coworker support, and physical demand scales) functioned differently between Chinese and Korean nurses.

However, to our knowledge, no study has examined cross-language DIF of the JCQ among European industrialized countries. While two international comparison studies of psychosocial job hazards[28] (Karasek et al., 2003, unpublished manuscript) used existing European JCQ datasets, they did not examine cross-language DIF of JCQ items beyond identifying the same factor structure in each country by exploratory factor analysis. Thus, the robustness of the cross-national or cross-regional mean comparisons of the JCQ scales in those studies remains in question.

However, another question follows: Whether or not the impact of DIF items on the scale level comparisons will be "substantial". It is a much more practical and important question since the JCQ international comparisons or most of epidemiologic studies with the JCQ have been done at the scale-level rather than at the item-level. In the aforementioned DIF study [27], despite many DIF items of the JCQ, there were no significant impacts of the DIF items on the scale-level comparisons. Such a result was also reported in other DIF studies [15, 29].

To address these two questions, we revisited one of the aforementioned international comparison studies (Karasek et al., 2003, unpublished manuscript) that used the European JCQ database from Belgium, France, Italy, The Netherlands, and Sweden of the Job Stress, Absenteeism, and Coronary Heart Disease European Cooperative Study (JACE Study) [30].

The objectives of this study were to assess the extent of cross-language DIF of the 27 JCQ items (Table 1) among six research centers (Table 2) of the JACE study and to test whether its effect on the scale-level mean comparisons is substantial. In addition, a judgment review on translation equivalence between the Flemish and Dutch JCQs was employed for an exploration of possible causes of DIF items statistically identified. The multi-central and multi-

Table 1 The 27 items of the job content questionnaire analyzed for the cross-language DIF analysis

Scales	Abbreviated items
Skill discretion (6 items)	Q3: "learn new things"
	Q4: "repetitive work"[a]
	Q5: "requires creative"
	Q7: "high skill level"
	Q9: "variety"
	Q11: "develop own abilities"
Decision authority (3 items)	Q6: "allows own decisions"
	Q8: "little decision freedom"[a]
	Q10: "lot of say"
Psychological demands (5 items)	Q19: "work fast"
	Q20: "work hard"
	Q22: "no excessive work"[a]
	Q23: "enough time"[a]
	Q26: "conflicting demands"[a]
Supervisor support (4 items)	Q48: "supervisor is concerned"
	Q49: "supervisor pays attention"
	Q51: "helpful supervisor"
	Q52: "supervisor good organizer"
Coworker support (4 items)	Q53: "coworkers competent"
	Q54: "coworker interest in me"
	Q56: "friendly coworkers"
	Q58: "coworkers helpful"
Physical demands (5 items)	Q21: "much physical effort"
	Q24: "lift heavy loads"
	Q25: "rapid physical activity"
	Q30: "awkward body positions"
	Q31: "awkward arm positions"

[a] Reversed in scale scoring and DIF analysis

language datasets made it possible to do various types of cross-language DIF analyses, comparing JCQ items in two different languages from the same country, two similar languages from two different countries, and two other languages from two different countries. Since this rich data set permits a large number of potential analyses, a new systematic methodology (Fig. 1) was devised and employed.

Table 2 Socio-demographic characteristics of the six populations for this study from the JACE-JCQ database

Centers	Brussels		Ghent		Lille		Milan[a]		Leiden		Gothenburg 95	
	M	W	M	W	M	W	M	W	M	W	M	W
Sample size	7,105	2,909	9,230	2,175	1,726	371	1,738	3,112	682	202	479	562
Sampling period (years)	1994 to 1998		1995 to 1998		1996 to 1997		1991 to 1997		1994 to 1996		1994 to 1995	
Age (years): Mean (SD)	45.9	44.3	45.9	44.4	43.0	42.0	44.4	42.8	42.8	43.7	47.5	47.4
	(6.0)	(5.7)	(6.0)	(5.6)	(5.8)	(5.9)	(6.7)	(5.5)	(5.4)	(5.4)	(7.3)	(7.0)
Education (years): Mean (SD)	12.5	12.0	12.5	12.1	11.9	13.9	10.4	9.4	10.4	11.2	12.4	12.5
	(3.3)	(2.8)	(3.1)	(2.8)	(2.8)	(3.5)	(2.9)	(3.1)	(5.5)	(5.6)	(3.9)	(3.4)
Translated JCQ (language)	Flemish[b]/French[c]		Flemish		French		Italian		Dutch		Swedish	

M men; *W* women

[a] Roughly 80% of Milan sample were used for this study

[b] Belgian-Dutch

[c] Belgian-French

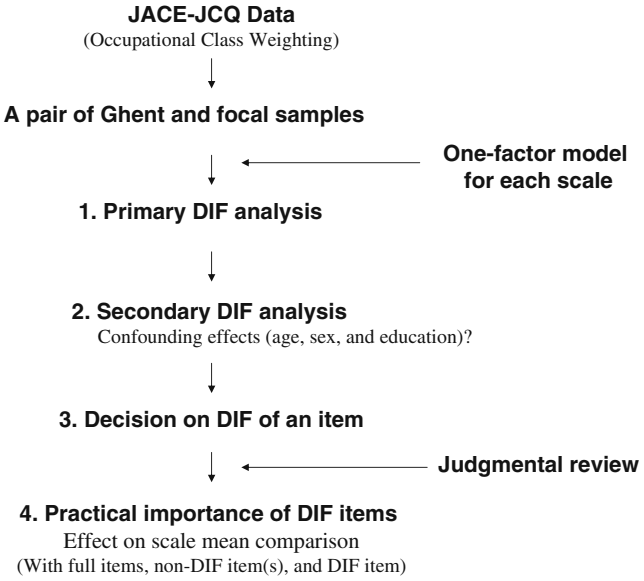

Fig. 1 DIF analysis procedure in the JACE-JCQ database

Materials and Methods

The JACE-JCQ Database

The JACE prospective epidemiology study [30] primarily used the original JCQ to measure perceived job stressors in five European countries from 1991 to 1998. However, different psychosocial questionnaires were used in the sample from Malmo, Sweden and two sub-sites (roughly 20%) of the Milan sample [31]. In addition, the Gothenburg 93 sample was comprised of only men aged 50 years. These datasets were therefore excluded from this study, leaving six populations from six research centers in five European countries: Belgium (Ghent and Brussels), France (Lille), Italy (Milan), The Netherlands (Leiden), and Sweden (Gothenburg 95). The number of participants in each study ranged from 884 to 11,405 (Table 2).

Generally, the six samples included broad distributions of detailed occupations identified according to the International Standard Classification of Occupations (ISCO) four-digit codes of the International Labor Office [32], except for the Milan sample. The Swedish center used a general population sample. The other centers recruited a more or less diverse private and public employee population from a broad range of organizations. The Milan sample included only public employees from six departments of the city administration. The ISCO one-digit compositions of each sample differed by sample and gender.

Most of their age ranges spanned 35 to 59. Years of education varied by site and gender and from a mean of 9 to 14 years. The education variable was missing in the Milan and Leiden samples, so it was replaced by an estimation of years based on educational level attained (Karasek et al., 2003, unpublished manuscript). For more detailed descriptions of the JACE study, refer to previous publications [28, 30, 33].

Samples and Items for Cross-Language DIF Analyses

In the five samples from Belgium (Brussels, Ghent), France (Lille), Italy (roughly 80% of the Milan sample), and The Netherlands (Leiden), 27 questions of the JCQ (Table 1) were used, while 22 questions were utilized in the Swedish sample (Table 3). All JCQ items used a four-Likert type response: *strongly disagree*, *disagree*, *agree*, and *strongly agree*. American English JCQ items were translated to Belgian-Dutch (Flemish), Belgian-French, French, Italian, Dutch, and Swedish and then back-translated to English to assess the semantic equivalence by each research center [28]. In the Belgium samples, both Belgian-Dutch (Flemish) and Belgian-French versions of the JCQ were administered: Ghent (Flemish speakers, 100%) and Brussels (Belgian-French speakers, 67.7% and Flemish speakers, 32.3%).

Only the English back-translation of the Flemish JCQ was available for this study. The items, Q7, Q19, Q20, and Q52 had been noted to have translation nuances in a posteriori review on the back-translation of the Flemish JCQ by the JCQ Center.

Weighting Samples

To prevent potential effects of different occupation compositions among the samples on DIF analyses, each sample for both men and women was weighted by the composition percentages of ISCO one-digit codes of the full JACE [33]. The weighting process generally resulted in considerable reduction of standard deviations of the JCQ scale means (27% for men and 46% for women, respectively, across all samples) [33].

DIF Analysis Procedures

First, exploratory factor analyses generally supported the assumption of one factor for each of the JCQ scales in all samples (Table 1). However, Q26 (from the psychological demands scale) had factor loadings of less than 0.30 in most of the samples [33]. Cronbach's alpha values of the JCQ scales, on average, ranged from 0.59 to 0.86 [33] (Fig. 1).

Second, cross-language DIF analyses were done between the reference and focal samples for each of the 27 JCQ items. The Ghent (Belgian-Dutch, Flemish) sample was chosen as the reference for other language samples

Table 3 The results of the primary and secondary DIF analyses of the 27 JCQ items between Ghent sample (reference) and each focal sample of the JACE-JCQ database

Scales	JCQ items	Brussels[a]	Lille		Milan		Leiden		Gothenburg 95		Percent of samples with DIF (primary analysis)	Percent of samples with DIF (secondary analysis)
		Primary	Primary	Secondary[b]	Primary	Secondary[b]	Primary	Secondary[b]	Primary	Secondary[c]		
Skill discretion	Q3	O	+	+	O	O	−	−	O	O	40	40
	Q4	O	−	−	−	−	−	O	O	O	60	40
	Q5	O	O	O	−	−	O	O	O	O	20	20
	Q7	O	−	−	+	+	+	+	+	O	80	60
	Q9	O	O	O	O	O	O	O	O	O	0	0
	Q11	+	+	+	+	+	+	+	+	O	100	80
Decision authority	Q6	O	O	O	−	O	O	O	O	+	20	20
	Q8	O	O	O	O	O	O	O	O	O	0	0
	Q10	O	O	O	+	+	O	O	O	O	20	20
Psychological demands	Q19	O	O	O	−	−	O	O	−	−	40	40
	Q20	O	O	O	+	+	+	+	O	O	40	40
	Q22	O	O	O	O	O	+	+	+	+	40	40
	Q23	O	−	−	−	−	O	O	O	O	40	40
	Q26	O	+	+	O	O	−	−	O	O	40	40
Supervisor support	Q48	−	−	−	O	−	O	O			50	75
	Q49	O	O	O	O	O	O	O			0	0
	Q51	O	+	+	−	−	−	−			75	75
	Q52	O	+	+	+	+	+	+			75	75
Coworker support	Q53	O	O	O	O	O	O	O	+	+	20	20
	Q54	O	O	O	O	O	O	O	+	+	20	20
	Q56	O	O	O	O	O	O	O	−	−	20	20
	Q58	−	−	−	O	O	−	−	−	−	80	80
Physical demands	Q21	O	O	O	−	−	O	O	−	O	40	20
	Q24	O	+	+	−	−	O	O	+	O	60	40
	Q25	O	O	O	O	O	−	−	O	O	20	20
	Q30	O	O	O	+	+	O	O	O	O	20	20
	Q31	O	−	−	+	+	O	O			50	50
% of DIF items:		11.1	44.4	44.4	55.6	55.6	40.7	37.0	45.5	27.3	39.2	36.2

O non-DIF item; (+) or (−) DIF item ('moderate to large DIF'—category C, see Bjorner et al. [15]; (+) over-endorsement by focal sample; (−) under-endorsement by focal sample

[a] Secondary analysis not done due to similar sample characteristics

[b] Controlled for sex and education

[c] Controlled for sex and age

(Brussels, Belgian-French/Flemish; Lille, French; Milan, Italian; Leiden, Dutch; Gothenburg 95, Swedish) to enable comparison of two linguistically similar languages (Dutch and Flemish) from two different countries, as well as two different languages from the same country (Belgian French vs. Belgian-Dutch). The Ghent sample also had the advantages of a large sample size, the second purpose of this study (impact of DIF items on the scale-level mean comparisons), availability of English back-translation, and documented psychometric validity [34]. Due to one missing item in the physical demand scale in the Gothenburg 95 sample, a four-item (Q21, Q24, Q25, and Q30) version of the physical demand scale was constructed for the DIF analyses.

Third, the partial gamma coefficient method [15, 17, 24, 35] was used for DIF statistics. The partial gamma coefficient is a variant of Kendall's τ, which is zero when two observations are as likely to be discordant as concordant given the conditional independence between item and variable of interest. Partial gamma coefficients were initially calculated at each score of a scale and finally combined across scale scores. To simplify the interpretation of cross-language DIF and detect the most pronounced differences, we chose the criterion, "moderate to large" DIF (category C) over "slight to moderate" DIF (category B) of Bjorner et al. [15]. Category C was defined as items with partial gamma outside the interval (−0.31 to 0.31) and its 95% confidence interval significantly outside the interval (−0.21 to 0.21); category A (no or negligible DIF) as items with partial gamma within the interval (−0.21~0.21) or its 95% confidence interval including zero; category B as items located between categories A and C.

Fourth, the impact of differential socio-demographic characteristics between Ghent and focal samples on the DIF analyses above (called "primary" DIF analyses hereafter) was examined. The primary DIF analyses were replicated (called "secondary" DIF analyses hereafter) after controlling for both sex and education between Ghent and Lille, between Ghent and Milan, and between Ghent and Leiden; both sex and age between Ghent and Gothenburg 95. There was no need for the secondary DIF analyses between Ghent and Brussels due to their similar sample characteristics. Education and age were both dichotomized (up to vs. greater than 12 years of education; up to vs. greater than 45 years old). The results of the secondary DIF analyses were conservatively preferred to those of the primary DIF analyses, considering the potential confounding effects of age, sex, and education.

Fifth, as an exploration of causes of DIF items (i.e., category C) statistically identified, the translation equivalence of the 27 items between the Flemish and Dutch JCQs was evaluated by two trilingual (English/Flemish/Dutch) researchers (authors EC and MB) who had not been involved in the translation process for either version or the DIF statistical analysis. They were asked independently (a) to evaluate conceptual non-equivalence (e.g., very fast vs. very hard) of each JCQ item between the two versions and (b) to report any differences in terms of missing or adding words (e.g., very fast vs. fast). They then were asked to come up with a final set of agreed evaluations through discussion, particularly on their initial, discrepant evaluations (on eight items). The final evaluation was compared with the result of the statistical DIF analysis between the two versions.

Sixth, the impact of identified DIF items on the mean comparisons of the JCQ scales between Ghent and each of the focal samples was examined in separate analyses for men and women. Three criteria were applied to judge the impact. The means of each full JCQ scale were compared. The comparison was then replicated with the reduced scale with non-DIF items in the secondary DIF analyses. If the two mean comparisons differed in terms of rank order of samples with statistical significance (alpha value=0.01), it was suspected that DIF items substantially affected the scale-level mean comparison. Then, it was finally considered if the mean comparison of the reduced scale "with" DIF items was similar to those with the full scale. For sensitivity test, an effect size measure, Cohen's d (the difference of means divided by the pooled standard deviation; 0.20, "small"; 0.50, "medium"; and 0.8, "large" [36]), was additionally employed.

Finally, the impact of identified DIF items on the mean comparisons of the JCQ scales among multi-language samples (i.e., Ghent, Brussels, Lille, Milan, Leiden, and Gothenburg 95) was also examined by sex. The means of the JCQ scales with the full items, the best non-DIF item(s) and the worst DIF item were compared with the same criteria as in the sixth procedure above. The multiple comparisons were undertaken with Student–Newman–Keuls test (alpha=0.001).

The SPSS (version 16.0) statistic program was used for all statistical analyses.[1]

[1] The following SPSS syntax was used for partial gamma coefficients.

CROSSTABS

variables=*var*x (minvalue,maxvalue), *var*y (minvalue, maxvalue), *var*z, (minvalue, maxvalue)

/tables=*var*x by *var*y by *var*z

/statistics=gamma

/format=avalue tables

Results

Cross-Cultural Comparisons of the DIF Items by Site

Fifty-one of the total tested 130 items (39.2%) appeared to be DIF items in the primary DIF analyses between Ghent and the focal samples. The percentages of DIF items were varied by center and JCQ scale (Table 3). They were high with the Milan sample (55.6%) and low with the Brussels sample (11.1%). The results of the secondary DIF analyses were very similar to those of the primary DIF analyses, although the number of DIF items substantially decreased in the secondary DIF analysis with Gothenburg 95 sample. The total number of DIF items slightly decreased to 47 of 130 items (36.2%) in the secondary DIF analyses.

The items, Q7 and Q11 of the skill discretion scale, Q48, Q51, and Q52 of the supervisor support scale, Q58 of the coworker support scale, and Q31 of the physical demand scale appeared to be DIF items in half or more of both the primary and secondary DIF analyses. All of the coworker support items in the Gothenburg 95 sample were DIF items in the secondary DIF analyses with Ghent sample (Table 3).

At the scale level, the decision authority scale had the most DIF-free items (on average, 86.7%) across the five primary DIF analyses, followed by coworker support (75.0%), physical demand (62.5%), psychological demands (60.0%), skill discretion (50.0%), and supervisor support (50.0%). In the case of the five secondary DIF analyses, the decision authority scale was also the best (on average, 86.7% DIF-free), followed by coworker support (75.0%), physical demand (75.0%), psychological demands (62.5%), skill discretion (60.0%), and supervisor support (56.3%). The skill discretion, psychological demands, and supervisor support scales were the most affected by DIF in both primary and secondary analyses.

Comparison Between Judgmental Reviews on Translation Equivalence and Statistical DIF Analyses

In the judgmental review, no translation problems were observed for the four items (i.e., Q7, Q19, Q20, and Q52) that had been noted to have translation nuances in the posteriori review on the back-translation of the Flemish JCQ. Instead, the review indicated some slight differences between the Flemish and English JCQs in items Q6, Q10, Q26, Q51, and Q52 (Table 4). For instance, the word “a lot” in item Q10 (English) was missed in the Flemish version.

There were various types of translation differences in the 14 JCQ items (Table 4): missing/adding a word (Q10, Q26, Q51, Q52, and Q54); different frequency-or-extent-related adverbs/adjectives (Q9 and Q6); translation nuance (Q4, Q11, Q53, and Q25); translation nuance plus missing/adding of a word (Q30, and Q31); and obvious conceptual difference (Q48).

The above groups of items were identified as category A, B, or C in the statistical DIF analyses, which reflects differential significance of translation differences in their respective item context.

In total, five out of ten DIF items (category C) in statistical analyses were associated with translation differences noted by the reviewers. The other DIF items were not associated with any translation differences: Q3, Q7, Q20, Q22, and Q58; they were all categorized as category C even if they were judged as highly translation equivalent in the independent review.

Impact of DIF Items on the Scale-Level Mean Comparisons with the Ghent Sample

The decision authority, psychological demands, and physical demands scales had no DIF items between the Ghent and Brussels samples. The means of skill discretion with both the full items and the non-DIF items were significantly higher in Ghent men than in Brussels men (Table 5). In contrast, the mean of skill discretion with the DIF item (Q11) was significantly higher in the Brussels sample than in the Ghent sample. Likewise, the mean comparisons of the other two JCQ scales between Ghent and Brussels by sex were not affected substantially by their DIF items. The differences of Cohen’s *d* values between the scale-mean comparisons with the full items and with the non-DIF items were less than 0.10.

Two of the 36 mean comparisons of the JCQ scales (that included at least one DIF item) between Ghent and the other focal samples for both men and women appeared to be substantially affected by DIF. The two were related to the lack of non-DIF items: coworker support between Ghent and Gothenburg 95 for both men and women. The differences of Cohen’s *d* values between the scale mean comparisons with the full items and with the non-DIF items were not greater than 0.20 (i.e., small [36]) in almost all of the other 34 comparisons. However, the differences of Cohen’s *d* values were between 0.20 and 0.50 (i.e., middle [36]) in the two comparisons: The mean differences of skill discretion between Ghent and Milan samples for both men and women were much smaller when the non-DIF items (Q3 and Q9) were used for the mean comparisons than when the full items were used.

Impact of DIF Items on the Scale-Level Mean Comparisons Among Multi-Language Samples

The rank-orders of the multi-language samples for skill discretion with the full items and the non-DIF item (Q9)

Table 4 Independent judgmental review on the translation comparability of the 27 JCQ items between the Flemish and Dutch JCQs

Scales	JCQ items	Judgmental reviews The Flemish JCQ : The Dutch JCQ (conceptual difference or any missing/adding words?)	Statistical DIF analysis (category of DIF items)
Skill discretion	Q3	–	C
	Q4	'Repetitief'(repetitive) : 'kortdurende' (short lasting)	B
	Q5	–	A
	Q7	–	C
	Q9	'Nogal' (quite some) : 'heel veel' (a lot of)	A
	Q11	'Bijzondere bekwaamheden' (special abilities) : 'vakbekwaamheid' (profession)	C
Decision authority	Q6	'Dikwijls' (often)[a] : 'veel' (many)	A
	Q8	–	A
	Q10	Missing word for 'a lot'[a] : 'veel' (a lot)	A
Psychological demands	Q19	–	B
	Q20	–	C
	Q22	–	C
	Q23	–	A
	Q26	Adding 'op het werk' (at work)[a] : no word for it	C
Supervisor support	Q48	'Ondergeschikten' (subordinates) : 'medewerkers' (co–workers)	A
	Q49	–	A
	Q51	'Mijn werk' (my work)[a] : 'het werk' (work in general)	C
	Q52	'Goed' (well) – work well together [a]: no word for it – work together	C
Coworker support	Q53	'Competente' (competent) and 'weten' (know) : 'goed (well)	A
	Q54	'Als person' (as an individual) : missing word for it	A
	Q56	–	B
	Q58	–	C
Physical demand	Q21	–	A
	Q24	–	A
	Q25	'Inspanning' (effort) : 'activiteiten' (activities)	C
	Q30	'Onnatuurlijke' (unnatural) : 'ongemakkelijk' (uncomfortable) 'vaak en gedurende lange periodes' (often AND during long periods) :'Vaak langdurig' (often during long periods)	B
	Q31	'Onnatuurlijke' (unnatural) : 'ongemakkelijk' (uncomfortable) no word for 'vaak' (often) : adding 'vaak'	B

[a] Slight differences with the English JCQ items

(−) no comments (high translation equivalence); *A* 'no or negligible DIF'; *B* 'slight to moderate DIF'; *C* 'moderate to large DIF'

were very similar to each other (Table 6). However, Milan had significantly higher skill discretion with the non-DIF item particularly in women: The rank of Milan women substantially changed from one of the lowest with the full items and the DIF item (Q11) to the third highest with the non-DIF item. The ranks of Leiden for decision authority with the non-DIF item (Q8) for both men and women were significantly higher, compared to those with the full items or DIF item.

There was no DIF-free item for psychological demands across the samples. In addition, the percentages of DIF items across the samples were 40% in all of the five psychological job demands items (see Table 3). Therefore, the two items, Q20 and Q22, were arbitrarily chosen as the best non-DIF items and one item, Q23, as the worst DIF item for the multiple sample comparisons. Milan women had one of the lowest psychological demand means with the full scale and the DIF item (Q23) but not with the non-DIF item (Q22) (Table 6). Swedish coworker support value was not comparable to other samples, and there was no substantial DIF impact case in the multi-sample mean comparison for physical demand.

Discussion

This study examined cross-language differences in the meaning of 27 JCQ items and the impact of those

Table 5 The mean comparisons of the skill discretion, supervisor support, and coworker support scales of the JCQ with the respective full items, non-DIF items, and DIF item between Ghent and Brussels samples

JCQ scales[a]	Items used for scales	Men		Cohen's *d*	*p* value	Women		Cohen's *d*	*p* value
		Brussels	Ghent			Brussels	Ghent		
Skill discretion	Full items=(Q3+Q4+Q5+Q7+Q9+Q11)×2	34.67	35.20	−0.08	.000	31.67	32.29	−0.09	.003
	Non-DIF items=(Q3+Q4+Q5+Q7+Q9)×2×(6/5)	34.74	35.81	−0.17	.000	31.80	32.87	−0.15	.000
	DIF item=(Q11)×2×6	34.10	32.09	0.22	.000	30.89	29.29	0.17	.000
Supervisor support	Full items=(Q48+Q49+Q51+Q52)	10.87	10.71	0.06	.000	10.83	10.56	0.10	.001
	Non-DIF items=(Q49+Q51+Q52)×(4/3)	10.97	10.64	0.13	.000	10.86	10.46	0.15	.000
	DIF item=(Q48)×4	10.55	10.90	−0.11	.000	10.80	10.86	−0.02	.581
Coworker support	Full items=(Q53+Q54+Q56+Q58)	12.15	12.02	0.07	.000	11.96	12.02	−0.03	.234
	Non-DIF items=(Q53+Q54+Q56)×(4/3)	12.20	12.00	0.11	.000	12.04	12.02	0.01	.761
	DIF item=(Q58)×4	11.96	12.07	−0.05	.003	11.68	12.00	−0.13	.000

[a] No 'moderate to large' cross-language DIF was found in the decision authority, psychological demands, or physical demands scales

differences on the scale mean values in a large dataset from five European countries. Despite the very similar factor structure among the samples, 36–39% of the total tested items showed cross-language DIF. The impacts of the DIF items on the mean comparisons of the JCQ scales among the six multi-language centers were non-trivial: underestimated skill discretion (Milan), underestimated decision authority (Leiden), underestimated psychological demands (Milan women), and incomparable coworker support (Gothenburg 95). Furthermore, a comparison of the JCQ translations into Flemish and Dutch suggested non-equivalence for one half of the DIF items. Cross-language differences, from translation or from cultural norms, at least among European languages, should be considered in any international comparative study using the JCQ scales.

Methodology of Cross-Language DIF Analysis

Item response theory (IRT) models and multi-group confirmatory factor analysis method are known to be the most advanced and sophisticated methods for DIF statistics [37, 38]. However, the applicability of IRT models highly depends on the sample size to obtain stable parameters. More importantly, its applicability to job and occupational analysis data has not yet been fully explored [39]. We think that the partial gamma coefficient method has advantages over multi-group confirmatory factor analysis in terms of simplicity, understandability, and applicability to wide ranges of sample sizes.

The procedure for cross-language DIF analysis in this study was methodologically robust (Fig. 1). Every step of the procedure was necessary and indeed contributed to reducing errors in DIF analyses. However, our DIF analyses might underestimate the extent of cross-language DIF items. First, we focused on "moderate to large" (Category C) DIF items, which was a realistic choice considering the multi-language DIF analyses of this study. Applying a stricter criterion (category B, "slight to moderate") [15] would have produced a higher number of DIF items. Second, only weighted partial gamma across scale scores were used for the criterion of cross-language DIF. However, the weighted partial gamma of an item may not reflect the DIF of the item at a specific range of scale scores, which means that DIF, at a specific range of scale scores, could be overlooked.

To do multi-language DIF analyses involving at least three languages imposed additional analytical difficulty. As the number of languages for comparison increases, more DIF items are likely to be found, and the probability of finding non-DIF items across the multi-languages decreases. It was inevitable to use the next best non-DIF item(s) across the samples in order to complete the mean comparisons for some JCQ scales (in case that there was no single DIF-free item across the samples). The three criteria of substantial impacts of DIF items on the mean comparisons of the JCQ scales were considered jointly in this study due to the following reasons. First, we think that either rank-order change or statistical significance change of the sample means of the JCQ scales is not perfect alone because the former tends to exaggerate trivial impacts and the latter depends on sample sizes. However, rank-order change needs to be more weighted in the small data, considering reduced power of statistical significance test. Second, we think that the third criterion (i.e., similarity of the mean comparisons of each of the JCQ scales with its full items and with its DIF items) is a necessary condition because a significant discrepancy of the mean comparisons between the full scale and the reduced scale with non-DIF items could occur for other reasons (e.g., multidimensionality of a scale).

Table 6 The mean comparisons of the JCQ scales with the respective full items, non-DIF items, and worst DIF item among the six samples of the JACE-JCQ database

JCQ scales	Used items for scales	Brussels	Ghent	Lille	Milan	Leiden	Goth.95	*p* value	Brussels	Ghent	Lille	Milan	Leiden	Goth.95	*p* value
		Men: means(rank, rank group based on statistical significance test)							Women: means(rank, rank group based on statistical significance test)						
Skill discretion	Full items	34.67 (5, II)	35.20 (3, II)	35.66 (2, II)	31.06 (6, III)	35.14 (4, II)	36.96 (1, I)	0.000	31.67 (4, II)	32.29 (3,II)	32.96 (2,II)	30.40 (6,III)	31.42 (5, II)	35.51 (1,I)	0.000
	Non-DIF item (Q9)	36.23 (5, II)	36.53 (3, II)	38.05 (1, I)	34.56 (6, III)	36.48 (4, II)	37.74 (2, I)	0.000	33.58 (5, II)	34.09 (4, II)	35.32 (2, I)	35.21 (3, II)	33.07 (6,II)	37.14 (1,I)	0.000
	DIF item (Q11)	34.10 (4, II)	32.09 (5, III)	36.46 (1, I)	30.88 (6, III)	35.11 (3, I)	35.52 (2, I)	0.000	30.89 (4, II)	29.29 (6, III)	34.95 (1,I)	30.02 (5, II)	31.84 (3,II)	34.12 (2, I)	0.000
Decision authority	Full items	34.90 (4, II)	34.96 (3, II)	36.46 (1, I)	30.38 (6, III)	34.78 (5, II)	36.29 (2, I)	0.000	32.89 (3,II)	32.68 (4,II)	34.24 (2, I)	31.32 (6,III)	31.54 (5,III)	34.57 (1, I)	0.000
	Non-DIF item (Q8)	33.54 (5, II)	34.54 (4, II)	35.70 (3, I)	29.60 (6, III)	35.74 (2, I)	36.33 (1, I)	0.000	33.19 (4,I)	33.16 (5, I)	34.31 (3, I)	31.82 (6,II)	34.55 (2,I)	35.32 (1, I)	0.000
	DIF item[a] (Q6)	37.09 (3, II)	36.84 (4, II)	37.91 (2, I)	30.80 (6, III)	36.19 (5,II)	39.26 (1, I)	0.000	33.67 (4,II)	33.78 (3, II)	35.68 (2,I)	30.12 (6, IV)	30.27 (5, IV)	36.87 (1, I)	0.000
Psychological demands	Full items	31.37 (4, III)	30.47 (6, III)	30.72 (5, III)	32.97 (1, I)	31.72 (3, II)	32.52 (2, I)	0.000	32.17 (4, I)	31.11 (6,II)	33.01 (1, I)	31.30 (5,II)	32.23 (3, I)	32.64 (2, I)	0.000
	Non-DIF item (Q20)	32.37 (1, I)	30.32 (4,II)	31.63 (3,I)	NA	NA	32.02 (2, I)	0.000	33.47 (2, I)	32.00 (4,II)	34.67 (1, I)	NA	NA	32.09 (3, II)	0.000
	Non-DIF item (Q22)	30.68 (2, II)	29.56 (3, II)	29.18 (4,III)	33.95 (1, I)	NA	NA	0.000	31.43 (2, I)	30.09 (4,II)	31.36 (3,I)	32.31 (1,I)	NA	NA	0.000
	DIF item[b] (Q23)	27.88 (5, II)	28.84 (3, II)	25.64 (6, III)	29.27 (2, II)	28.55 (4, II)	31.26 (1, I)	0.000	29.05 (4, II)	29.54 (2,II)	27.92 (5,II)	26.80 (6, III)	29.15 (3, II)	31.39 (1, I)	0.000

NA not applicable due to 'moderate to large DIF' (category C, Bjorner et al. [15]); *Goth* Gothenburg

[a] Only in Goth. 95 (see Table 3)

[b] Only in Lille and Milan (see Table 3)

Possible Causes of Statistical DIF Items

We cannot determine specific causes of cross-language DIF items statistically identified in this study. However, some possible sources can be discussed, and several conspicuous patterns of the DIF items across the JACE samples deserve to be discussed.

The most distinct source of the cross-language DIF items in the JACE database seems to be translation-related difference. A half of the statistical DIF items between the Flemish and Dutch JCQs were associated with translation differences, ranging from a simple missing/adding word to obvious translation non-equivalence. The proportion (50%) was not unusual, compared to those (27–44%) in other cross-language DIF studies [40, 41]. It is also understandable, considering the fact that there was no pre-designed protocol for addressing translation equivalence across the research centers in the JACE study (Houtman et al. 1998).

However, it needs to be remembered that the other half of the statistical DIF items were not related to any translation differences in the judgmental review. In addition, the proportion of DIF items was much smaller in the analysis between two different language samples (Ghent and Brussels) from the same country than in the analyses between two Dutch speaking samples (Ghent and Leiden). All these imply that a national-level culture [42], interacting with structures and functions of institutions, might play a role as a source of DIF of the JCQ items in the JACE database.

The items highly vulnerable to cross-language DIF in the JACE database were Q7, Q11, Q48, Q51, Q52, Q58, and Q31. The skill discretion, psychological demands, and supervisor support scales were the most affected. This is consistent with the finding of Karasek et al. [26], with respect to inconsistent meanings of psychological demands items across the populations from industrialized countries. In addition, this study suggests that other JCQ items (particularly, supervisor support items) may be also differently understood among European countries. One reason for that may be that "demands and social support reflect to a great extent local work site conditions and individual perception" ([43], p. 18). Furthermore, the scale-level differential impact of DIF items might provide a clue for relatively higher heterogeneous associations of psychological job demands and social support at work with common mental health across European countries, compared to those of decision authority [44]. The items that are prone to DIF need to be considered both for improving the quality of the existing translated versions of the JCQ and exploring unique cultural characteristics ("emic" approach, see Peng et al. [9]) among the European countries in the future. In addition, their vulnerability to cross-language DIF needs to be considered seriously in the future version of the JCQ (JCQ 2.0, http://www.jcqcenter.org).

To reduce cross-language DIF of the JCQ in the future, it will be desirable to employ a stricter translation process as confirmed in the case of the Flemish JCQ: The translation and back-translation procedure were less sensitive to quality of translation than the independent review, which is consistent with the previous studies [7, 19, 20]. Useful techniques also include quantitative DIF analyses and qualitative interview (e.g., see [45]).

International Comparison of Psychosocial Job Hazards using the Existing JCQ Data

This study suggests that the previous international mean comparison using the JACE-JCQ database (Karasek et al., 2003, unpublished manuscript) needs to be carefully reviewed, considering the DIF impacts on the scale-level mean comparisons identified in this study. It would be wise to use only non-DIF JCQ items for more accurate international comparisons with the JACE-JCQ datasets in the future.

Lastly, we emphasize that this study was undertaken with the JCQ database from the five European countries sharing relatively similar cultures. Thus, the measurement equivalence test of the global JCQ database from European, North American, Asian, and Latin American countries with significantly different cultures still remains to be tested in the future.

Acknowledgements The authors thank the JACE study group [M. Kornitzer, Chairman (Brussels); G. de Backer (Ghent); M. Romon-Rousseau/C. Boulenguez (Lille); I. Houtman (Leiden); L. Wilhelmsen (Gothenburg); P-O Ostergren (Malmo); M. Ferrario (Milan); S. Sans (Barcelona)] for support on this project.

References

1. Karasek RA, Gordon G, Pietrokovsky C, Frese M, Pieper C, Schwartz J, et al. Job content questionnaire and user's guide. Los Angeles: University of Southern California; 1985.
2. Hurrell JJ Jr, Nelson DL, Simmons B. Measuring job stressors and strains: where we have been, where we are, and where we need to go. J Occup Health Psychol. 1998;3:322–55.
3. Lansbergis P, Theorell T. Self-reported questionnaires. In: Schnall PL, Belkić K, Lansbergis P, Baker D, editors. Occupational medicine: State of the art reviews-the workplace and cardiovascular disease, vol. 15, no. 1. Philadelphia: Hanley & Belfus; 2000. p. 163–71.
4. Sauter SL, Brightwell WS, Colligan MJ, Hurrell JJ Jr, Katz TM, LeGrande DE, et al. The changing organization of work and the safety and health of working people: Knowledge gaps and research directions (DHHS publication no. 2002-116). Cincinnati: National Institute for Occupational Safety and Health; 2002.
5. World Health Organization (WHO). The world health report 2002: reducing risks, promoting healthy life. Geneva: WHO; 2002.

6. Cavusgil ST, Das A. Methodological issues in empirical cross-cultural research: a survey of the management literature and a framework. Manag Int Rev. 1997;37:71–96.
7. Hambleton RK. Issues, designs, and technical guidelines of adapting tests into multiple languages and cultures. In: Hambleton RK, Merenda PF, Spielberger CD, editors. Adapting educational and psychological tests for cross-cultural assessment. Mahwah: Erlbaum; 2005. p. 3–38.
8. Harkness J. In pursuit of quality: Issues for cross-national survey research. Int J Soc Res Methodol. 1999;2(2):125–40.
9. Peng TK, Peterson MF, Shyi YP. Quantitative methods in cross-national management research: Trends and equivalence issues. J Organ Behav. 1991;12:87–107.
10. Schaffer B, Riordan CM. A review of cross-cultural methodologies for organizational research: a best-practices approach. Organ Res Methods. 2003;6(2):169–215.
11. Van de Vijver FJR, Poortinga YH. Conceptual and methodological issues in adapting tests. In: Hambleton RK, Merenda PF, Spielberger CD, editors. Adapting educational and psychological tests for cross-cultural assessment. Mahwah: Erlbaum; 2005. p. 39–63.
12. Riordan CM, Vandenberg RJ. A central question in cross-cultural research: do employees of different cultures interpret work-related measures in an equivalent manner? J Manage. 1994;20(3):643–71.
13. Angolf WH. Use of difficulty and discrimination indices for detecting item bias. In: Berk RA, editor. Handbook of methods for detecting test bias. Baltimore: The Johns Hopkins University; 1982. p. 96–116.
14. Angolf WH. Perspectives on differential item functioning methodology. In: Holland PW, Wainer H, editors. Differential item functioning. Hillsdale: Erlbaum; 1993. p. 3–29.
15. 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):1189–202.
16. Camilli G, Shepard LA. MMSS: Methods for identifying biased test items. Thousand Oaks: Sage; 1994.
17. Ørhede E, Kreiner S. Item bias in indices measuring psychosocial work environment and health. Scand J Work Environ Health. 2000;26(3):263–72.
18. Bullinger M, Alonso J, Apolone G, Leplège A, Sullivan M, Wood-Dauphinee S, et al. Translating health status questionnaires and evaluating their quality: the IQOLA project approach. J Clin Epidemiol. 1998;51(11):913–23.
19. Bjorner JB, Thunedborg K, Kristensen TS, Modvig J, Bech P. The Danish SF-36 health survey: translation and preliminary validity studies. J Clin Epidemiol. 1998;51:991–99.
20. Hunt S, McKenna S. Cross-cultural comparability of quality of measures. Br J Med Econ. 1992;4:17–23.
21. Keller SD, Ware JE, Gandek B, Aaronson NK, Alonso J, Apolone G, et al. Testing the equivalence of translations of widely used response choice labels: results from the IQOLA project. J Clin Epidemiol. 1998;51(11):933–44.
22. Wagner AK, Gandek B, Aaronson NK, Acquadro C, Alonso J, Apolone G, et al. Cross-cultural comparisons of the content of SF-36 translations across 10 countries: results from the IQOLA project. J Clin Epidemiol. 1998;51(11):925–32.
23. Ellis BB, Kimmel HD. Identification of unique cultural response patterns by means of item response theory. J Appl Psychol. 1992;77(2):177–84.
24. Petersen MA, Groenvold M, Bjorner JB, Aaronson N, Conroy T, Cull A, et al. Use of differential item functioning analysis to assess the equivalence of translation of a questionnaire. Qual Life Res. 2003;12:373–85.
25. Kristensen TS. Job stress and cardiovascular disease: a theoretic critical review. J Occup Health Psychol. 1996;1(3):246–60.
26. Karasek RA, Brisson C, Kawakami N, Houtman I, Bongers P, Amick B. The job content questionnaire (JCQ): An instrument for internationally comparative assessment of psychosocial job characteristics. J Occup Health Psychol. 1998;3:322–55.
27. Choi B, Cho SL, Jian L, Yang W, Jhun HJ, Xu Z et al. (2006). A differential item functioning test for the cross-national quantitative comparison of psychosocial job hazards between Chinese and Korean nurses. Paper submitted to the 9th International Congress of Behavioral Medicine, Bangkok, Thailand.
28. de Smet P, Sans S, Dramaix M, Boulenguez C, de Backer G, Ferrario M, et al. Gender and regional differences in perceived job stress across Europe. Eur J Pub Health. 2005;15(5):536–45.
29. Stark S, Chernyshenko OS, Drasgow F. Examining the effects of differential item (functioning and differential) test functioning on selection decisions: when are statistically significant effects practically important? J Appl Psychol. 2005;89(3):497–508.
30. Houtman I, Kornitzer M, de Smet P, Koyunko R, de Barker G, Pelfrene E, et al. Job stress, absenteeism and coronary heart disease European cooperative study (the JACE study): design of a multicentre prospective study. Eur J Pub Health. 1999;9:52–7.
31. Karasek RA, Choi B, Ostergren PO, Ferrario M, de Smet P. Testing two methods to create comparable scale scores between the Job Content Questionnaire (JCQ) and JCQ-like questionnaires, in the European JACE study. Int J Behav Med. 2007;14(4):189–201.
32. International Labor Office. International standard classification of occupations: ISCO-88. Geneva: International Labor Office; 1990.
33. Choi B (2006) Methodological and theoretical issues in cross-national comparative studies of psychosocial job hazards: from questionnaire items to social class. Dissertation, University of Massachusetts, Lowell, MA
34. Pelfrene E, Vlerick P, Mak RP, de Smet P, Kornitzer M, de Barker G. Scale reliability and validity of the Karasek 'job demand-control-support' model in the Belstress study. Work Stress. 2001;15:297–313.
35. Kreiner S. Validation of index scales for analysis of survey data: the symptom index. In: Dean K, editor. Population health research: liking theory and methods. Thousand Oaks: Sage; 1993. p. 116–44.
36. Cohen J. A power primer. Psychol Bull. 1992;112(1):115–59.
37. Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. Psychol Bull. 1993;114(3):552–66.
38. Meade AW, Lautenschlager GJ. A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. Organ Res Methods. 2004;7(4):361–88.
39. Harvey RJ (2005) Applicability of binary IRT models to job analysis data. In A. Meade (Chair), Applications of IRT for measurement in organizations. Paper presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL. http://harvey.psyc.vt.edu/Documents/HarveySIOP2003.pdf. Accessed 2 Jan 2005
40. Ercikan K. Translation effects in international assessments. Int J Educ Res. 1998;29:543–53.
41. Ercikan K. Disentangling sources of differential item functioning in multilanguage assessments. Int J Testing. 2002;2(3&4):199–215.
42. Hofstede G. Cultural consequences: comparing values, behaviors, institutions, and organizations across nations, vol. 2. Thousand Oaks, CA: Sage; 2001. p. 209–372.
43. Theorell T, Karasek RA. Current issues relating to psychosocial job strain and cardiovascular disease research. J Occup Health Psychol. 1996;1(1):9–26.
44. Standsfeld S, Candy B. Psychosocial work environment and mental health—a meta-analytic review. Scand J Work Environ Health. 2006;32(6):443–62.
45. Mallinson S. Listening to respondents: a qualitative assessment of the Short-Form 36 Health Status Questionnaire. Soc Sci Med. 2002;54:11–21.