International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 18(1): 58–67 (2009)
Published online 4 February 2009 in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/mpr.277

Application of item response theory to achieve cross-cultural comparability of occupational stress measurement

AKIZUMI TSUTSUMI,[1] NOBORU IWATA,[2] NAOTAKA WATANABE,[3] JAN DE JONGE,[4] HYNEK PIKHART,[5] JUAN ANTONIO FERNÁNDEZ-LÓPEZ,[6] LIYING XU,[7] RICHARD PETER,[8] ANDERS KNUTSSON,[9] ISABELLE NIEDHAMMER,[10,11] NORITO KAWAKAMI[12] & JOHANNES SIEGRIST[13]

1 Occupational Health Training Center, University of Occupational and Environmental Health, Kitakyushu, Japan
2 Department of Clinical Psychology, Hiroshima International University, Japan
3 Graduate School of Business Administration, Keio University, Japan
4 Department of Technology Management, Eindhoven University of Technology, The Netherlands
5 International Institute for Society and Health, Department of Epidemiology and Public Health, University College London, UK
6 Health's Center of Riosa, Health Service of Asturias' Principality (SESPA), Riosa, Spain
7 Clinical Trial Consultant Company, Shanghai, China
8 Department of Epidemiology, University of Ulm, Germany
9 Department of Health Science, Mid Sweden University, Sweden
10 INSERM, U687-IFR69, France
11 UCD School of Public Health & Population Science, University College Dublin, Ireland
12 Department of Mental Health, Tokyo University Graduate School of Medicine, Japan
13 Institut für Medizinische Soziologie, Heinrich Heine Universität Düsseldorf, Germany

Key words
cross-cultural comparison, differential item functioning, differential test functioning, item response theory, stress

Correspondence
Akizumi Tsutsumi, University of Occupational and Environmental Health, Occupational Health Training Center, 1-1 Iseigaoka, Yahatanishi-ku, Kitakyushu 807-8555, Japan.
Telephone (+81) 93 691 7167
Fax (+81) 93 692 4590
Email: tsutsumi@med.uoeh-u.ac.jp

Received 10 September 2007;
revised 18 April 2008;
accepted 28 May 2008

Abstract

Our objective was to examine cross-cultural comparability of standard scales of the Effort–Reward Imbalance occupational stress scales by item response theory (IRT) analyses. Data were from 20,256 Japanese employees, 1464 Dutch nurses and nurses' aides, 2128 representative employees from post-communist countries, 963 Swedish representative employees, 421 Chinese female employees, 10,175 employees of the French national gas and electric company and 734 Spanish railroad employees, sanitary personnel and telephone operators. The IRT likelihood ratio model was used for differential item functioning (DIF) and differential test functioning (DTF) analyses. Despite the existence of DIF, most comparisons did not show discernible differences in the relations between Effort–Reward total score and level of the underlying trait across cultural groups. In the case that DTF was suspected, excluding an item with significant DIF improved the comparability. The full cross-cultural comparability of Effort–Reward Imbalance scores can be achieved with the help of IRT analysis.

Introduction

Management of adverse psychosocial working conditions is important for mental health in the current industrialized world (Marmot *et al.*, 2006). Economic globalization has contributed to an unprecedented degree of internationalization of companies and their staff where employees with different cultural backgrounds are collaborating or involved in transnational division of labor. This situation increasingly requires cross-culturally validated and standardized evaluations of a stressful psychosocial work environment in the global world of work.

At the conceptual level, these evaluations require a theoretical basis that identifies the most crucial (i.e. health-adverse) components within the complex realities of working life. At the methodological level, established psychometric properties of the scales measuring these components must be met, including a test of their cross-cultural comparability. Several conceptualizations of a stressful psychosocial work environment were developed (Antoniou and Cooper, 2005), but only a few proved their utility in terms of predicting incident mental disorders in prospective epidemiological studies. Among these, the Demand–Control model (Karasek and Theorell, 1990) and the Effort–Reward Imbalance model (Siegrist, 1996) deserve special attention. Their explanatory power has been reviewed at different occasions (Marmot *et al.*, 2006; Stansfeld and Candy, 2006; Tsutsumi and Kawakami, 2004; Van der Doef and Maes, 1999; van Vegchel *et al.*, 2005). Moreover, the psychometric properties of the scales measuring the models' key components have repeatedly been demonstrated in different countries (de Jonge *et al.*, 2008; Karasek *et al.*, 1998; Kawakami *et al.*, 1995; Li *et al.*, 2005; Niedhammer *et al.*, 2000, 2004; Siegrist *et al.*, 2004; Tsutsumi *et al.*, 2001). However, little information is available on their standardization for cross-cultural comparisons.

Item response theory (IRT) provides a good approach towards this end (Bjorner *et al.*, 1998; Raczek *et al.*, 1998). IRT is a psychometric theory that represents mathematical functions which relate person and item parameters to the probability of the responses. IRT also provides a basis for estimating parameters, ascertaining how well data fits a model, and investigating the psychometric properties of assessments (Hambleton *et al.*, 1991). It is also a reliable tool for identifying differential item functioning (DIF) or so-called item bias (Holland and Wainer, 1993). DIF refers to the observation that an item displays different statistical properties in different group settings, after controlling for differences in the latent traits of the groups (Angoff, 1993). Existence of DIF may indicate that the validity of comparisons is restricted, but the presence of one or more items exhibiting DIF should not prevent the scaling of individuals on a common metric at the level of total scores (differential test functioning; DTF) (Reise *et al.*, 1993). For example, summing items may cancel out or amplify their bias (Cooke *et al.*, 2001). DTF can be examined by plotting test characteristics curves (Lord, 1980). The test characteristics curves indicate how the association between the latent trait and a change in Effort–Reward scores varies across cultures.

The present paper applies IRT to the scales measuring the Effort–Reward Imbalance model of occupational stress and aims to explore whether the scales are cross-culturally comparable. In the context of IRT, DIF can be evaluated by examining whether a particular item parameter differs between the groups (Holland and Wainer, 1993; Peng *et al.*, 1991). In this paper, we compare the threshold parameters, which indicate the severity of an item response, of the Effort–Reward scale items between Japan (reference group) and other focal groups such as those from European and Asian countries.

Method

Study population

The study population comprised 20,256 Japanese employees from 20 work sites of 12 different occupations, such as dental technicians, manufacturing workers, nurses, hospital employees, software engineers, white-collar workers of businesses or the service sector, and employees of a cooperative society and a prefectural government (Tsutsumi, 2004), 1464 Dutch nurses and nurse's aides, 963 Swedish representative employees (the WOLF-Norrland Study) (Peter *et al.*, 1998), 2128 representative employees of post-communist countries (Poland, Czech Republic, Lithuania, and Hungary) (Pikhart *et al.*, 2001), 421 Chinese working women (Xu *et al.*, 2004), 10,175 employees of the French national gas and electric company (the GAZEL cohort) (Goldberg *et al.*, 2001), and 734 Spanish railroad employees, sanitary personnel, and telephone operators (Table 1). In all studied samples, the Effort–Reward Imbalance has been shown to be associated with several health outcomes.

The Effort–Reward Imbalance model questionnaire

The Effort–Reward Imbalance model assumes that 'Effort at work' (e.g. working overtime or time pressure) is spent as part of a contract based on the norm of social reciprocity where 'Rewards' are provided in terms of money, esteem, and career opportunities including job security.

Table 1 Characteristics of study population

Cultural groups	Data source	*n*	Women (%)	Mean age	Standard deviation	Range
Japan	Various occupations[a]	20256	44	39.7	11.0	17–75
The Netherlands	Care givers, nurse and nurse's aide	1464	88	39.5	9.6	16–69
France	National electric and gas company	10175	29	51.0	2.9	44–59
Sweden	Several companies representing different sectors in the northern region of Sweden	963	23	53.3	10.0	33–76
Post-communist countries	Representative community sample	2846	49	44.3	9.3	18–68
Spain	National railway staff, sanitary personnel, and Tele-operators	734	64	34.8	9.3	18–63
China	Four worksites in Beijing	421	100	14.2	10.2	0–39[b]

[a] Derived from 20 work sites of 12 different occupations including dental technicians, manufacturing workers, nurses, hospital employees, software engineers, white-collar workers of businesses or the service sector, and employees of a cooperative society and a prefectural government.
[b] Career years.

The model posits that work contracts often fail to provide a symmetric exchange, and that such 'high cost-low gain' conditions are relatively frequent in the modern economy (e.g. due to limited alternative choices in the labor market). Recurrent experiences of Effort–Reward Imbalance at work elicit strong negative emotions and stressful responses within the employee with adverse long-term effects on health (Siegrist, 2005). Finally, it is assumed that employees characterized by a motivational pattern of excessive job-related commitment and a high need for approval (i.e. overcommitment) will respond with more strain reactions to an Effort–Reward Imbalance, in comparison with less overcommitted people. To assess the personal component, a third unidimensional scale containing six items was developed. In this paper, however, we focus on the situation-specific components Effort and Reward that are most relevant for cross-cultural comparisons. Personal components such as overcommitment are more sensitive to individual aspects than these two work-related components.

Data were collected using the standardized questionnaire measuring Effort–Reward Imbalance (Siegrist *et al.*, 2004). Effort measures relevant features of a demanding daily work environment experienced by the employee (six Likert-scaled items). The Reward scale consists of 11 items covering the three aspects of financial, esteem-related and status-related rewards. On the scales Effort and Reward, the respondents are asked about whether stressful environmental conditions exist. If they agree, they are then asked to indicate the level of distress on a four-point scale, which ranges from 'very distressed' to 'not at all distressed.' The questionnaire has been globally developed through back-translation procedures, and the basic psychometric properties have been confirmed in each country, including invariance of factor structure across cultures using confirmatory factor analysis (de Jonge *et al.*, 2008; Macías Robles *et al.*, 2003; Niedhammer *et al.*, 2000, 2004; Siegrist *et al.*, 2004; Tsutsumi *et al.*, 2001; van Vegchel *et al.*, 2002; Xu *et al.*, 2004). It should be noted that one reward item (Reward 1) was not used for the post-communist countries, so this item was excluded from the analysis of this sample (see Appendix, Table A1). In all groups under study, the first principal factors accounted for more than 30% of variance of the two scales, thus justifying the application of the IRT model. The levels of Cronbach alpha for each scale were acceptable among all the groups (from 0.67 to 0.84 for the Effort scale, and from 0.79 to 0.89 for the Reward scale, respectively).

Statistical methods

The Japanese group, composed of diverse occupations, was designated as the reference group and the other cultural groups were referred to as focal groups. Item responses were coded as binary according to the manual (Siegrist and Peter, 1997). Scores were established by separating the answers to each item – the worst two categories versus the rest – and adding them together. The rating procedure was defined as follows: 1 = the worst two categories of intense distress and 0 = other. The reasons why we employed the dichotomized data are: (1) the

Effort–Reward Imbalance scale items were originally coded as binary to create the stress index, (2) it was easy to graphically display the results; displays produced by IRT models for categorical data would be too complex to clearly visualize the item characteristics, and (3) it was difficult to obtain acceptable model fit statistics for all the languages particularly in the comparisons among multicultural languages such as ours. We also considered it important to avoid complex discussion on selecting appropriate polytomous IRT models; simple binary models would provide enough information for practically comparable scale development.

For DIF analysis, the IRT likelihood ratio model was employed using BILOG-MG software (version 3.0) (du Toit, 2003). First, to conduct model-fit tests for the estimated item response functions, we used the value '–2 times the log of the likelihood function'; this value leads to the statistic G^2. If G^2 for the full invariance model was significantly greater than that for the baseline model, we could conclude that at least one item must contain DIF (Thissen *et al.*, 1993). Second, we inferred the item parameters by the marginal maximum likelihood method, and adopted a difference between thresholds over 0.3 as the criterion indicating the existing of DIF (Thissen *et al.*, 1988). Third, to examine DTF, we plotted test characteristics curves for ratings from the referent (Japan) versus those from each focal group (Lord, 1980). In the case that metric invariance was not supported, i.e. there was a large discrepancy between test characteristics curves, we re-examined DTF after excluding an item with significant DIF from the scale (Holland and Wainer, 1993).

To verify the results, we conducted three subanalyses. First, since it would make more sense to compare each focal group with the respective occupational group from the Japanese sample, we extracted the nurses and hospital staff (n = 1959 and 868, respectively) from the Japanese data, then redid the DIF analysis between the sub-sample and the Dutch sample. Second, since the high imbalance in sample sizes for the Japanese reference group and each focal group might affect the results, we randomly extracted a Japanese sub-sample of almost equal size with the focal group and then redid the DIF analyses. Third, since the study samples were inhomogeneous regarding the gender and age distribution, both variables were modeled as covariates in the IRT model. For this purpose, we estimated such extended IRT models in M*plus* (Muthén and Muthén, 1998–2007). We used gender and age (16–29, 30–39, 40–49, 50–75 years) as covariates in both a multiple group model where group is country and in a MIMIC model where country is a covariate along with gender and age. If we obtain a modification index for the item parameter, it is either fixed or constrained to be equal to another parameter. A modification index gives the expected drop in chi-square if the parameter in question is freely estimated. Any large modification index indicates that freeing the parameter or removing the equality constraint could result in better model fit, that is, the parameters for a particular item differs between groups (existing DIF). For simplicity and to avoid multiple comparisons, we made comparisons between the Japanese sample and the Dutch sample in the second and third cases, too.

Results

Comparisons of the log likelihood of the fit of the DIF and non-DIF models indicated significantly better fit of the DIF model ($p < 0.01$ for all comparisons; Table 2). We

Table 2 Values of G^2 at convergences in DIF and non-DIF models and the differences[a] (in parentheses)

	The Netherlands	France	Sweden	Post-communists countries	Spain	China
Effort						
Non-DIF	82316	102674	70593	94302	76302	72915
DIF	73281	102569	69772	80566	68771	66201
	(9035)	(105)	(821)	(13736)	(7531)	(6714)
Reward						
Non-DIF	117623	176094	111744	131035	110414	105600
DIF	84376	133220	83654	90991	83956	79659
	(33247)	(42874)	(28090)	(40044)[b]	(24458)	(25941)

[a] Distributed as chi-square on six degrees of freedom for Effort scale and 11 degrees of freedom for Reward scale.
[b] Distributed as chi-square on 10 degrees of freedom.

could expect that there is at least one item which displays DIF among each pair of cultural groups both in Effort and Reward scales.

Among the six Effort items, one to four items were found to display DIF when the Japanese group was compared with each focal group. As for the 11 Reward items, two to five items were found to display DIF (Table 3). Relatively larger numbers of DIF were found for the Effort scale than in the Reward scale, particularly in the comparisons between the Netherlands, France or Sweden and Japan. In contrast, items displaying DIF were fewer in Spain and China than in other countries.

However, presence of the few items exhibiting DIF did not necessarily affect metric invariance at the level of total score: aggregation across items with cross-cultural DIF appeared to result in cancellation of DIF (Figures 1 and 2). The cross-cultural differences were marked for the scores, where large numbers of DIF existed, in particular, for the Effort scores in the comparisons between the Netherlands or France and Japan.

Removal of items with significant DIF from the scale would lead to better comparability of the scores. As an illustration, take the case of the Netherlands and Japan. The Netherlands is one of the earliest countries where work sharing has been successfully introduced. Among the industrialized countries, the Dutch workers enjoy the shortest work time, whereas Japanese workers have the longest one (International Labour Organization, 2004). Thus, an Effort-item of 'pressure to work overtime' (Effort 4) may have different meanings between these two countries. In addition, this item appears to reflect a personal, rather than a situational, effort factor for Dutch workers (de Jonge *et al.*, 2003). Removing Effort 4 reduced the number of items that displayed DIF from four items out of six in the original Effort scale to one item out of five in the modified scale. The test characteristics curves made by the remaining five items indicate substantially improved metric invariance of total scores (Figure 3).

Comparison between the Japanese sub-sample of nurses and hospital staff and the Dutch sample (similar

Table 3 IRT slope parameters (*a*) and adjusted threshold parameters (*b*) of the Effort–Reward scale items between Japan and other cultural groups

	Japan versus the Netherlands			Japan versus France			Japan versus Sweden			Japan versus post-communist countries			Japan versus Spain			Japan versus China		
Item	*a*	b_r	b_f	*a*	b_r	b_f	*a*	b_r	b_f	*a*	b_r	b_f	*a*	b_r	b_f	*a*	b_r	b_f
Effort 1	4.3	1.3	1.0	5.2	1.3	1.7*	4.7	1.3	1.0	4.3	1.3	1.3	4.6	1.3	1.1	4.7	1.3	1.2
Effort 2	2.6	1.8	1.5*	2.9	1.8	2.0	2.8	1.7	1.3*	2.7	1.8	1.2*	2.9	1.7	1.4	2.9	1.7	1.6
Effort 3	2.5	1.3	1.7*	2.7	1.4	1.8*	2.6	1.3	1.7*	2.6	1.3	1.7*	2.7	1.3	1.5	2.7	1.3	1.6*
Effort 4	3.0	1.6	2.2*	3.0	1.6	1.4	3.1	1.6	1.7	3.0	1.6	1.6	3.2	1.5	1.9*	3.3	1.5	1.5
Effort 5	3.0	1.5	1.2*	2.0	1.8	0.3*	3.0	1.5	1.8*	3.0	1.5	1.7	3.2	1.5	1.6	3.3	1.5	1.4
Effort 6	3.7	1.3	1.2	3.0	1.4	2.1*	3.8	1.3	1.2	3.8	1.3	1.3	4.0	1.3	1.1	4.0	1.3	1.2
Reward 1	4.3	1.6	1.6	4.0	1.6	1.4	4.2	1.6	2.1*	–	–	–	4.1	1.6	1.8	4.2	1.6	1.8
Reward 2	4.3	1.8	2.1	3.3	1.9	2.1	4.3	1.9	1.5*	3.6	1.9	2.2	4.1	1.9	2.2	4.2	1.9	2.3*
Reward 3	3.1	2.0	1.7	2.8	2.0	1.7*	3.0	2.0	1.9	2.9	2.0	1.9	2.9	2.0	1.8	3.0	2.0	2.1
Reward 4	4.3	1.7	2.1*	4.6	1.6	1.7	4.1	1.7	2.0*	3.7	1.7	1.7	4.1	1.7	2.0*	4.2	1.7	1.6
Reward 5	2.3	1.6	1.4	2.1	1.6	1.5	2.3	1.6	1.6	2.3	1.6	1.6	2.3	1.6	1.6	2.3	1.6	1.4
Reward 6	3.6	1.7	1.5	3.8	1.7	1.3*	3.4	1.8	1.5*	3.6	1.8	1.8	3.4	1.8	1.5	3.6	1.8	1.6
Reward 7	2.4	1.9	2.3*	2.3	2.0	2.6*	2.3	2.0	2.0	2.4	2.0	1.6*	2.3	2.0	1.8	2.3	2.0	1.2*
Reward 8	2.8	1.8	2.0	2.8	1.8	2.0	2.7	1.8	2.0	2.8	1.8	2.3*	2.7	1.8	1.9	2.8	1.8	2.1
Reward 9	5.5	1.6	1.4	6.6	1.6	1.5	4.9	1.7	1.0*	5.2	1.6	1.6	5.5	1.6	1.6	5.4	1.6	1.6
Reward 10	3.3	1.6	1.6	4.1	1.5	1.4	3.2	1.6	1.6	3.5	1.5	1.6	3.2	1.6	1.7	3.2	1.6	1.8
Reward 11	1.9	1.6	1.2*	2.2	1.5	1.7	1.9	1.6	1.9	1.9	1.6	1.1*	1.9	1.6	1.3*	1.9	1.6	1.7

Note: *a* indicates slope parameters which is assumed to be equal over the compared groups in the analyses; b_r indicates adjusted threshold parameters of referent group (Japan); b_f indicates adjusted threshold parameters of focal groups. Asterisk indicates existence of DIF.

Int. J. Methods Psychiatr. Res. 18(1): 58–67 (2009). DOI: 10.1002/mpr

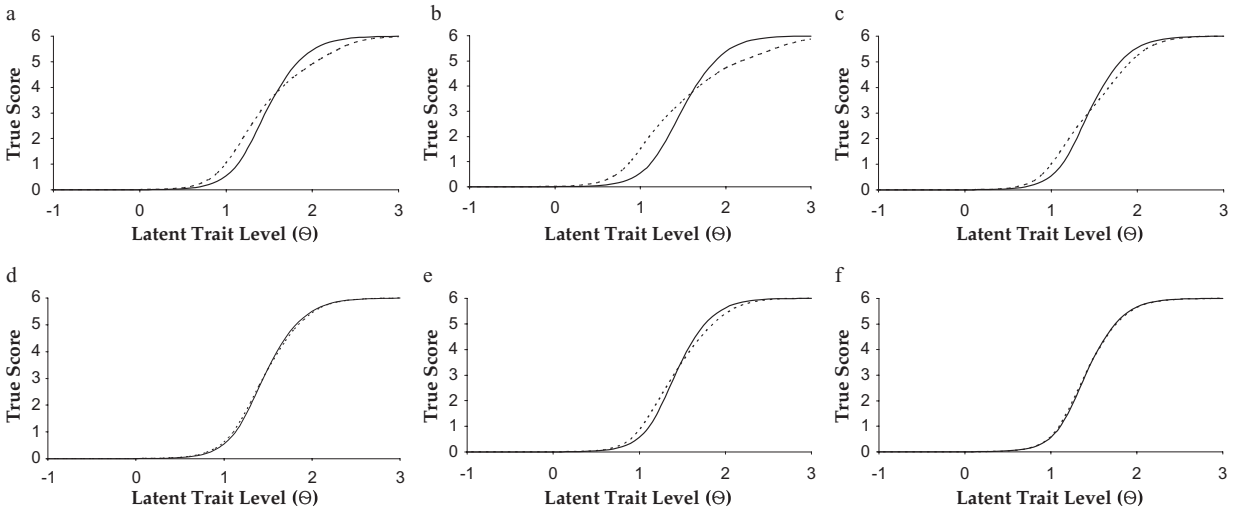

Figure 1 Test characteristics curves for Effort scores: reference (Japan; solid line) versus focal groups (a) the Netherlands, (b) France, (c) Sweden, (d) post-communist countries, (e) Spain, and (f) China (dotted line).

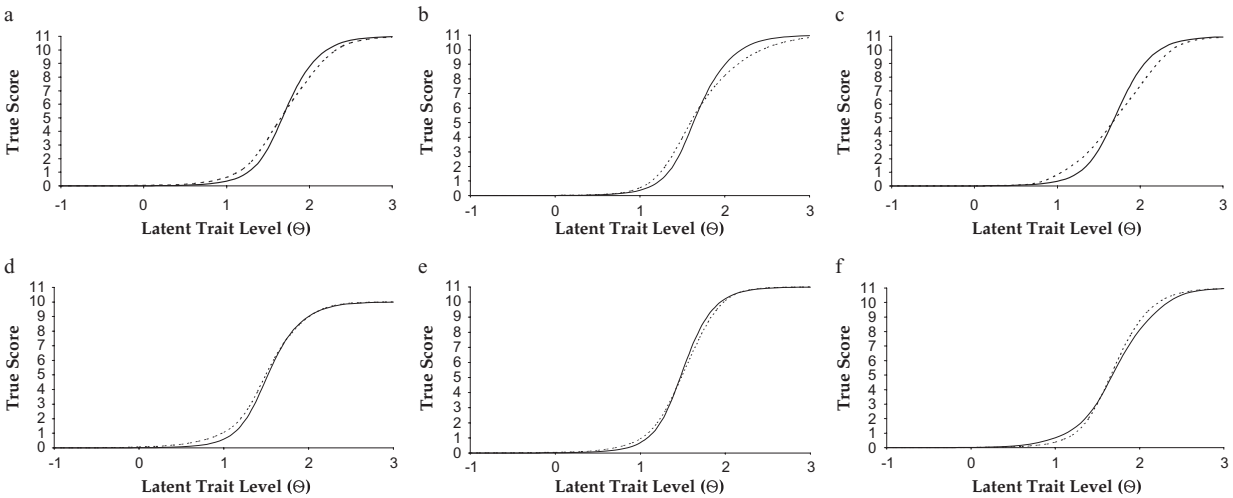

Figure 2 Test characteristics curves for Reward scores: reference (Japan; solid line) versus focal groups (a) the Netherlands, (b) France, (c) Sweden, (d) post-communist countries, (e) Spain, and (f) China (dotted line).

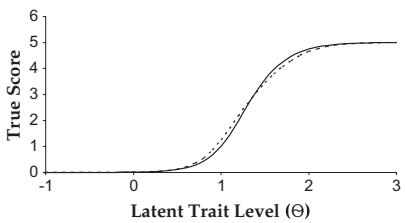

Figure 3 Test characteristics curves for Effort scores after excluding item 'pressure to overtime work': Japan (solid line) versus the Netherlands (dotted line).

occupations) did not necessarily provide better comparability of the Effort scores than the comparison between the total Japanese sample and the Dutch sample. With slightly different items from the main analysis with the total Japanese sample, four items were found to display DIF for Effort scale, and there was a large discrepancy between the test characteristics curves. Although removing Effort 4 that displayed the greatest magnitude of DIF did not reduce the numbers of items displaying DIF according to the defined criterion, the differences in threshold parameters between the sub-sample and the

Dutch sample substantially decreased and produced almost identical test characteristics curves. Second, we randomly extracted a sub-sample from the total Japanese sample which was almost the same sample size as the Dutch sample ($n = 1426$) and then redid the DIF analyses. The results provided almost the same pattern as the results of the main analysis. Removal of Effort 4, which again displayed the greatest magnitude of DIF, produced almost identical test characteristics curves. Lastly, we modeled gender and age as covariates in the IRT models. The analyses revealed the greatest modification index for Effort 4 followed by Effort 3, and removing the equality constraint of Effort 4 between the countries got rid of any significant modification indices for the remained items. The gender- and age-adjusted findings indicated that Effort 4 was the only item exhibiting significant DIF between the countries. Thus, all the subanalyses confirmed the results of main analysis (data not shown).

Discussion

By applying IRT analysis to internationally used standardized scales measuring the Effort–Reward Imbalance model of a stressful psychosocial work environment, we observed a number of DIF among the items between Japan and other cultural groups. But the differences in item functioning did not necessarily restrict a comparative evaluation of the measurements across these countries (DTF), as summing items canceled out their item differences in the majority of comparisons. Even in the case of suspected DTF, excluding a single item with significant DIF improved the comparability.

Scores on the Effort scale obtained in Japan are not directly comparable with those obtained in the Netherlands. However, by excluding a single item from the scale a substantial reduction both in the number of DIF and in the metric variance of total scores can be reached. Choice of the items would need empirical tests based on the actual data between the cultural groups concerned. Removal of the item displaying a greater magnitude of DIF is a practical solution (Holland and Wainer, 1993) and a socially regulated working system and empirical information regarding the psychometric property of the scale would support the decision. In the case of our illustration, the excluded item was characterized as the obvious difference in work time arrangements as well as in the unique psychometric properties of the item in an earlier Dutch study (de Jonge *et al.*, 2003). Evidently, this item is not well suited to representing the common construct of extrinsic Effort between the two cultural groups, and including this item may distort the construct and reduce comparability of respective measurements across cultural groups.

In exchange for good comparability, excluding items may suppress useful information in different societal contexts. For example, the item 'pressure to work overtime' is an important item to measure work overload particularly among the Japanese workers. IRT offers a solution for this 'etic-emic' dilemma (Peng *et al.*, 1991). Constraining core (etic) items to have identical parameters across groups (anchoring) ensures that responses are underpinned by a latent trait with a common metric. For the core items, the same set of parameters is assumed to apply to both groups. This ensures that trait levels and item parameters for the non-core items are estimated on the same scale and thus are directly comparable. Future study would include creating compatible measures of a certain construct with culture-specific (emic) items.

Items exhibiting DIF more often belonged to the Effort scale than to the Reward scale. This fact may be attributable to well-known difficulties of measuring 'demanding factors' at work, a construct that is composed of different dimensions (Kristensen *et al.*, 2004). This might be particularly critical if such diverse populations and occupational groups are compared, as is the case with the current study (Steenland *et al.*, 1997). In addition, our analyses explored the similarity and dissimilarity of the responses among countries. The numbers of DIF were relatively fewer between Japan and Spain or China. A cross-cultural comparison with respect to the Demand–Control model indicated similar findings: the levels of reported stress were relatively similar between employees in Japan and in southern European countries (de Smet *et al.*, 2005; Kawakami *et al.*, 2004).

Our study has as its strengths a large data set in addition to diversity and variance in the sample, allowing for reliable parameter estimates. One of the advantages of IRT is the parameter invariance; each item characteristic is expressed by a few parameters that are estimated independent of the sample distribution (Hambleton and Cook, 1977; Lord and Novick, 1968). This advantage allows IRT not to require random sampling to estimate the parameters for comparison, and is of special interest in the framework of cross-cultural validation of psychosocial questionnaires (Embretson and Reise, 2000). We made the best use of the sample and the advantage of IRT. The three subanalyses confirmed the robustness of the results; the differences of occupations, sample size and inhomogeneity of demographic characteristics appeared not to affect the results of comparisons between the cultural groups.

The results of this cross-cultural analysis indicate caution in directly comparing the scores of the Effort–Reward Imbalance scales in different cultural groups. In particular this applies to the Effort scale. As was shown the decision of removing single items with significant DIF may substantially improve comparability of measurements, thus posing a trade-off between a better comparability and a more comprehensive operational measurement of an underlying construct. Comparability of occupational stress measurement is an increasing need in an era of globalized working life. Although the limits of this statistical approach need to be taken into account, application of IRT to scales measuring a stressful psychosocial work environment offers a promising perspective.

Acknowledgments

Thanks are due Drs Keiichi Eguchi, Masahiro Irie, Tsuyoshi Kato, Yuri Kawano, Akiko Miki, Akinori Nakata, Yuko Odagiri, Yumiko Oya, Akihito Shimazu, Teruichi Shimomitsu, Katsutoshi Tanaka, and Naoko Toyoda for offering their valuable data for the standardization of the Japanese version of the Effort–Reward Imbalance questionnaire.

This study was partly supported by Health and Labour Sciences Research Grants (Research on Occupational Safety and Health; Research number H17-Rodo-2).

Declaration of interests statement

The authors declare that they have no competing interests.

References

Angoff W.H. (1993) Perspectives on differential item functioning methodology. In: *Differential Item Functioning* (eds Holland P.W., Wainer H.), pp. 3–23, Lawrence Erlbaum.

Antoniou A.-S.G., Cooper C.L. (2005) *Research Companion to Organizational Health Psychology (New Horizons in Management)*, Edward Elgar Publishing.

Bjorner J., Kreiner S., Ware J., Damsgaard M., Bech P. (1998) Differential item functioning in the Danish translation of the SF-36. *Journal of Clinical Epidemiology*, **51**, 1189–1202, DOI: 10.1016/S0895-4356(98)00111-5

Cooke D.J., Kosson D.S., Michie C. (2001) Psychopathy and ethnicity: structural, item, and test generalizability of the Psychopathy Checklist-Revised (PCL-R) in Caucasian and African American participants. *Psychological Assessment*, **13**, 531–542.

de Jonge J., van der Linden S., Schaufeli W., Peter R., Siegrist J. (2003) Validity and reliability of the Dutch ERI scales: preliminary findings. ESF meeting. Düsseldorf.

de Jonge J., van der Linden S., Schaufeli W., Peter R., Siegrist J. (2008) Factorial invariance and stability of the Effort–Reward Imbalance Scales: a longitudinal analysis of two samples with different time-lags. *International Journal of Behavioral Medicine*, **15**, 62–72.

de Smet P., Sans S., Dramaix M., Boulenguez C., de Backer G., Ferrario M., Cesana G., Houtman I., Isacsson S.O., Kittel F., Östergren P.O., Peres I., Pelfrene E., Romon M., Rosengren A., Wilhelmsen L., Kornitzer M. (2005) Gender and regional differences in perceived job stress across Europe. *European Journal of Public Health*, **15**, 536–545, DOI: 10.1093/eurpub/cki028

du Toit M. (2003) IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT, Scientific Software International.

Embretson S.E., Reise S.P. (2000) *Item Response Theory for Psychologists*, Lawrence Erlbaum.

Goldberg M., Chastang J.F., Leclerc A., Zins M., Bonenfat S., Bugel I., Kaniewski N., Schmaus A., Niedhammer I., Piciotti M., Chevalier A., Godard C., Imbernon E. (2001) Socioeconomic, demographic, occupational, and health factors associated with participation in a long-term epidemiologic survey: a prospective study of the French GAZEL cohort and its target population. *American Journal of Epidemiology*, **154**, 373–384.

Hambleton R.K., Cook L.L. (1977) Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement*, **24**, 75–96.

Hambleton R.K., Swaminathan H., Rogers H.J. (1991) *Fundamentals of Item Response Theory (Measurement Methods for the Social Sciences)*, Sage Publications.

Holland P.W., Wainer H. (1993) *Differential Item Functioning*, Lawrence Erlbaum.

International Labour Organization (2004) *Working Time and Workers' Preferences in Industrialized Countries: Finding the Balance*, Routledge.

Karasek R., Brisson C., Kawakami N., Houtman I., Bongers P., Amick B. (1998) The Job Content Questionnaire (JCQ): an instrument for internationally comparative assessments of psychosocial job characteristics. *Journal of Occupational Health Psychology*, **3**, 322–355.

Karasek R., Theorell T. (1990) *Healthy Work: Stress, Productivity, and the Reconstruction of Working Life*, Basic Books.

Kawakami N., Kobayashi F., Araki S., Haratani T., Furui H. (1995) Assessment of job stress dimensions based on the job demands-control model of employees of telecommunication and electric power companies in Japan: reliability and validity of the Japanese version of the Job Content Questionnaire. *International Journal of Behavioral Medicine*, **2**, 358–375.

Kawakami N., Haratani T., Kobayashi F., Ishizaki M., Hayashi T., Fujita O., Aizawa Y., Miyazaki S., Hiro H., Masumoto T., Hashimoto S., Araki S. (2004) Occupational class and exposure to job stressors among employed

men and women in Japan. *Journal of Epidemiology*, **14**, 204–211.

Kristensen T.S., Bjorner J.B., Christensen K.B., Borg V. (2004) The distinction between work pace and working hours in the measurement of quantitative demands at work. *Work & Stress*, **18**, 305–322.

Li J., Yang W., Cheng Y., Siegrist J., Cho S.-I. (2005) Effort–reward imbalance at work and job dissatisfaction in Chinese healthcare workers: a validation study. *International Archives of Occupational and Environmental Health*, **78**, 198–204, DOI: 10.1007/s00420-004-0581-7

Lord F.M. (1980) *Applications of Item Response Theory to Practical Testing Problems*, Lawrence Erlbaum.

Lord F.M., Novick M.R. (1968) *Statistical Theories of Mental Test Scores*, Addison-Wesley.

Macías Robles M.D., Fernández-López J.A., Hernández-Mejía R., Cueto-Espinar A., Rancaño I., Siegrist J. (2003) [Measuring psychosocial stress at work in Spanish hospital's personnel. Psychometric properties of the Spanish version of Effort–Reward Imbalance model]. *Medicina Clinica*, **120**, 652–657.

Marmot M., Siegrist J., Theorell T. (2006) Health and the psychosocial environment at work. In: *Social Determinants of Health* (eds Marmot M., Wilkinson R.G.), pp. 97–130, Oxford University Press.

Muthén L.K., Muthén B.O. (1998–2007) *Mplus User's Guide*, fifth edition, Muthén & Muthén.

Niedhammer I., Tek M.-L., Starke D., Siegrist J. (2004) Effort–reward imbalance model and self-reported health: cross-sectional and prospective findings from the GAZEL cohort. *Social Science & Medicine*, **58**, 1531–1541, DOI: 10.1016/S0277-9536(03)00346-0

Niedhammer I., Siegrist J., Landre M.F., Goldberg M., Leclerc A. (2000) Étude des qualités psychométriques de la version française du modèle du Déséquilibre Efforts [Psychometric properties of the French version of the Effort–Reward Imbalance model]. *Revue d'Épidémiologie et de Santé Publique*, **48**, 419–437.

Peng T.K., Peterson M.F., Shyi Y. (1991) Quantitative methods in cross-national management research: trends and equivalence issues. *Journal of Organizational Behavior*, **12**, 87–107.

Peter R., Alfredsson L., Hammar N., Siegrist J., Theorell T., Westerholm P. (1998) High effort, low reward, and cardiovascular risk factors in employed Swedish men and women: baseline results from the WOLF study. *Journal of Epidemiology and Community Health*, **52**, 540–547.

Pikhart H., Bobak M., Siegrist J., Pajak A., Rywik S., Kyshegyi J., Gostautas A., Skodova Z., Marmot M. (2001) Psychosocial work characteristics and self rated health in four post-communist countries. *Journal of Epidemiology and Community Health*, **55**, 624–630.

Raczek A., Ware J., Bjorner J., Gandek B., Haley S., Aaronson N., Apolone G., Bech P., Brazier J., Bullinger M., Sullivan M. (1998) Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: results from the IQOLA Project. International Quality of Life Assessment. *Journal of Clinical Epidemiology*, **51**, 1203–1214, DOI: 10.1016/S0895-4356(98)00112-7

Reise S.P., Widaman K.F., Pugh R.H. (1993) Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological Bulletin*, **114**, 552–566.

Siegrist J. (1996) Adverse health effects of high-effort/low-reward conditions. *Journal of Occupational Health Psychology*, **1**, 27–41.

Siegrist J. (2005) Social reciprocity and health: new scientific evidence and policy implications. *Psychoneuroendocrinology*, **30**, 1033–1038, DOI: 10.1016/j.psyneuen.2005.03.017

Siegrist J., Peter R. (1997) *Measuring Effort–Reward Imbalance at Work: Guidelines*, Düsseldorf: Heinrich Heine Universität Düsseldorf.

Siegrist J., Starke D., Chandola T., Godin I., Marmot M., Niedhammer I., Peter R. (2004) The measurement of Effort–Reward Imbalance at work: European comparisons. *Social Science & Medicine*, **58**, 1483–1499, DOI: 10.1016/S0277-9536(03)00351-4

Stansfeld S., Candy B. (2006) Psychosocial work environment and mental health – a meta-analytic review. *Scandinavian Journal of Work, Environment & Health*, **32**, 443–462.

Steenland K., Johnson J., Nowlin S. (1997) A follow-up study of job strain and heart disease among males in the NHANES1 population. *American Journal of Industrial Medicine*, **31**, 256–260, DOI: 10.1002/(SICI)1097-0274(199702)31:2

Thissen D., Steinberg L., Wainer H. (1988) Use of item response theory in the study of group differences in trace lines. In: *Test Validity* (eds Wainer H., Braun I.), pp. 147–169, Lawrence Erlbaum.

Thissen D., Steinberg L., Wainer H. (1993) Detection of differential item functioning using the parameters of item response models. In: *Differential Item Functioning* (eds Holland P.W., Wainer H.), pp. 67–113, Lawrence Erlbaum.

Tsutsumi A. (2004) *Development and Application of the Japanese Version of Effort–Reward Imbalance Questionnaire*. Report of Research Project, Grand in Aid for Scientific Research (C), 2001~2004, Okayama University School of Medicine and Dentistry, Hygiene & Preventive Medicine.

Tsutsumi A., Kawakami N. (2004) A review of empirical studies on the model of effort–reward imbalance at work: reducing occupational stress by implementing a new theory. *Social Science & Medicine*, **59**, 2335–2359, DOI: 10.1016/j.socscimed.2004.03.030

Tsutsumi A., Ishitake T., Peter R., Siegrist J., Matoba T. (2001) The Japanese version of the Effort–Reward

Imbalance questionnaire: a study in dental technicians. *Work & Stress*, **15**, 86–96.

Van der Doef M., Maes S. (1999) The job demand-control(-support) model and psychological well-being: a review of 20 years of empirical research. *Work & Stress*, **13**, 87–114.

van Vegchel N., de Jonge J., Bakker A.B., Schaufeli W.B. (2002) Testing global and specific indicators of rewards in the Effort–Reward Imbalance model: does it make any difference? *European Journal of Work and Organizational Psychology*, **11**, 403–421.

van Vegchel N., de Jonge J., Bosma H., Schaufeli W.B. (2005) Reviewing the effort–reward imbalance model: drawing up the balance of 45 empirical studies. *Social Science & Medicine*, **60**: 1117–1131, DOI: 10.1016/j.socscimed.2004.06.043

Xu L., Siegrist J., Cao W., Li L., Tomlinson B., Chan J. (2004) Measuring job stress and family stress in Chinese working women: a validation study focusing on blood pressure and psychosomatic symptoms. *Women & Health*, **39**, 31–46.

Appendix

Table A1 The Effort–Reward scales

Effort scale	
Effort 1	I have constant time pressure due to a heavy work load.
Effort 2	I have many interruptions and disturbances in my job.
Effort 3	I have a lot of responsibility in my job.
Effort 4	I am often pressured to work overtime.
Effort 5	My job is physically demanding.
Effort 6	Over the past few years, my job has become more and more demanding.
Reward scale	
Reward 1	I receive the respect I deserve from my superiors.
Reward 2	I receive the respect I deserve from my colleagues.
Reward 3	I experience adequate support in difficult situations.
Reward 4	I am treated unfairly at work.
Reward 5	I have experienced or I expect to experience an undesirable change in my work situation.
Reward 6	My job promotion prospects are poor.
Reward 7	My job security is poor.
Reward 8	My current occupational position adequately reflects my education and training.
Reward 9	Considering all my efforts and achievements, I receive the respect and prestige I deserve at work.
Reward 10	Considering all my efforts and achievements, my work prospects are adequate.
Reward 11	Considering all my efforts and achievements, my salary / income is adequate.