Qual Life Res (2010) 19:219–224
DOI 10.1007/s11136-009-9577-9

Cross-national comparability of the WHOQOL-BREF: A measurement invariance approach

Peter Theuns · Joeri Hofmans · Mehrdad Mazaheri · Frederik Van Acker · Jan L. Bernheim

Accepted: 19 December 2009 / Published online: 20 January 2010

Abstract
Purpose To evaluate whether the WHOQOL-BREF measures the QOL construct in the same way across nations.
Methods Students from Flanders, Belgium and Iran completed the WHOQOL-BREF as part of a larger Quality of Life questionnaire. Their responses were compared using a multi-group confirmatory factor analysis.
Results In general, the QOL construct appears rather similar in both cultures; however, participants from both countries seem to respond differently to particular items of the WHOQOL-BREF. Especially for the physical and psychological domain, this is problematic, because none of their indicators works in the same way across samples.
Conclusions Notwithstanding some limitations of this study, it must be concluded that the WHOQOL-BREF should only be used with great caution in cross-national comparisons.

Keywords Quality-of-life construct · Measurement invariance · Intercultural comparability · Life domains

P. Theuns (✉) · F. Van Acker
Vakgroep Experimentele en Toegepaste Psychologie, Faculty of Psychology and Education Sciences, EXTO, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium
e-mail: Peter.Theuns@vub.ac.be

J. Hofmans (✉)
Research Group of Quantitative Psychology and Individual Differences, Katholieke Universiteit Leuven, Tiensestraat 102, 3000 Leuven, Belgium
e-mail: Joeri.Hofmans@psy.kuleuven.be

M. Mazaheri
Department of Psychology, University of Sistan & Baluchestan, Zahedan, Iran
e-mail: mazaheri@hamoon.usb.ac.ir

Present Address:
F. Van Acker
Faculty of Psychology, Open Universiteit Nederland, PO Box 2960, 6401 DL Heerlen, The Netherlands
e-mail: Frederik.VanAcker@ou.nl

J. L. Bernheim
Human Ecology, Faculty of Medicine and Pharmacy, MEKO, Vrije Universiteit Brussel, Laarbeeklaan 103, 1090 Brussel, Belgium
e-mail: jan.bernheim@vub.ac.be

Abbreviations

QOL	Quality of life
WHOQOL-BREF	World health organization quality of life assessment instrument
ACSA	Anamnestic comparative self assessment
PWI	Personal wellbeing index
RMSEA	Root mean square error of approximation
CFI	Comparative fit index

Introduction

The sensitivity of Quality of life (QOL) measures to cultural differences is an important problem in cross-national research [1–4]. It has been found that financial satisfaction [5, 6], satisfaction with education, material wealth, home life, health [7, 8], satisfaction with esteem needs [1, 5, 7], and number of friends [8] are differentially related to overall life satisfaction in different populations.

Accordingly, it can be questioned whether different determinants contribute similarly to QOL in different populations.

The WHOQOL-BREF [9] was introduced as an instrument to "*... assess variation in QOL across different cultures, to compare subgroups within the same culture and to measure change across time in response to change in life circumstances*" [9 on page 2].

Although the WHOQOL-BREF was found to be a sound and valid instrument for measuring QOL [10], it remains an empirical question whether it measures the same construct in different populations. Moreover, even if this question is answered affirmatively, a subsequent question is whether this instrument measures those constructs in the same way, so that average scores can be compared reliably. The present study will address both questions by evaluating the cross-national measurement invariance of the WHOQOL-BREF.

Method

Participants

Iranian and Dutch-speaking Belgian psychology students (i.e., from the Flemish part of Belgium) were invited to participate in a study on happiness. Participants consented to fill out a questionnaire on their QOL after being informed that their personal data would be treated confidentially. The samples are described in Table 1.

Materials and procedure

The WHOQOL-BREF [9] was included in the questionnaire. This instrument consists of 26 items measuring four domains considered to contribute to overall QOL: psychological, physical, social, and environmental well-being.

Table 1 Characteristics of the samples

	Flanders ($N = 389$)	Iran ($N = 150$)
Gender		
Male	97	16
Female	287	134
No info	5	
Age		
Mean (SD)	22.60 (5.61)	18.93 (2.30)

All participants are students of psychology at a university (Flanders, Belgium) or a school for higher education (Iran). Participants with incomplete data (75 Belgian, 5 Iranian) were not included in the analyses

Twenty-four items measure the respective domains –these are called "facets" items–and the other two—which are called "benchmark" items–measure general well-being.

Data analysis

Measurement invariance is tested using multi-group confirmatory factor analysis (CFA), which is not robust against violating the assumption of continuous factor indicators [11, 12]. Consequently, because the items of the WHOQOL-BREF are ordinal, an estimation method that explicitly takes this ordinality into account was used (more precisely, the Weighted Least Squares estimator with a mean- and variance-adjusted Chi-square (WLSMV) on a polychoric correlation matrix along with the Delta parameterization in Mplus [11, 13]). Apart from selecting an appropriate estimation method, conducting a multi-group CFA with ordinal indicators requires a specialized approach. In particular, our data analysis follows the procedure as outlined by Millsap and Yun-Tein [12], which consists of a series of increasingly restrictive nested tests according to the following sequence: (1) testing invariance of the factor structure, (2) step 1 + testing invariance of the factor loadings, and (3) step 2 + testing invariance of the thresholds. A fourth test, that is, step 3 + a test on residual variances is not performed as it is not required for the meaningful comparison of means [14].

Overall, model fit is tested by evaluating different fit indices simultaneously. The χ^2 is reported, because it is a common test in CFA [15], although it is known to be overly sensitive to minor violations of the model and to be severely affected by sample size [15]. In addition, the root mean square error of approximation (*RMSEA*) and the comparative fit index (*CFI*) are reported. As a rule of thumb, *RMSEA* <.05 indicates close fit, values between .05 and .08 suggest reasonable fit, and values greater than .10 suggest poor fit [15]. For the *CFI*, values above .90 can be expected for a reasonably good fitting model [15, 16]. Apart from these global fit indices, incremental fit indices are used when comparing consecutive models, because the tests in this paper are nested. According to Cheung and Rensvold [17], a ΔCFI larger than .01 would indicate a serious reduction in fit. Apart from ΔCFI, $\Delta\chi^2$ is also reported, although it is known to suffer from the same problems as the ordinary χ^2 (i.e., high sensitivity to minor violations of the model). It should be noted that the rules of thumb we present (for the global as well as for the incremental fit indices) should not be considered as if they were golden rules. We concur with a large number of scientists that these fit indices are just one of the many pieces of information on which a decision about a model should be based [see 18].

Table 2 Thresholds, fit parameters, and drop in model fit resulting from equating thresholds for each consecutive WHOQOL-BREF item

Threshold			Model fit and caused fit deterioration in partial metric invariance test				
Item	Flanders, Belgium	Iran	*CFI*	Δ*CFI*	$\Delta\chi^2$	*df*	*sign.*
3	**.164/.950/1.408/1.923**		**.907**	**.000**	**7.241**	**2**	**.027**
4	.605/1.120/1.698/2.278		.907	.000	2.087	3	.555
5*	**−2.048/−1.556/−.291/1.158**	**−2.048/−1.556/−.931/−.299**	**.907**	**.000**	**45.734**	**2**	**<.001**
6*	−1.629/−1.086/−.244/1.063	−1.629/−1.265/−.709/−.304	.906	.001	56.970	3	<.001
7*	−1.680/−.896/.311/1.833	−1.680/−1.204/−.641/−.013	.906	.001	51.799	3	<.001
8	**−2.189/−1.457/−.507/1.067**		**.907**	**.000**	**7.453**	**2**	**.024**
9	−2.282/−1.439/−.198/1.382		.907	.000	6.972	3	.073
10*	−2.306/−1.051/−.029/1.158	−2.306/−2.157/−1.549/−1.254	.893	.014	182.537	3	<.001
11*	−1.972/−.997/−.113/1.097	−1.972/−2.075/−.950/−.362	.904	.003	76.392	3	<.001
12	−2.093/−1.209/−.042/.881		.907	.000	7.120	3	.068
13	−2.663/−1.587/−.289/1.062		.907	.000	2.849	3	.416
14	−1.787/−.572/.393/1.352		.907	.000	4.056	3	.255
15*	−3.002/−1.948/−.544/1.146	−3.002/−2.273/−1.576/−.952	.902	.005	797.499	3	<.001
16*	−1.659/−.701/−.172/1.086	−1.659/−1.466/−1.267/−.792	.893	.014	962.052	3	<.001
17*	−2.383/−1.310/−.530/1.325	−2.383/−1.991/−1.645/−1.160	.894	.013	172.597	3	<.001
18*	−2.167/−1.184/−.386/1.171	−2.167/−1.859/−1.556/−.998	.895	.012	162.358	3	<.001
19*	−2.179/−1.280/−.457/1.295	−2.179/−1.826/−1.226/−.472	.906	.001	99.351	3	<.001
20	**−2.120/−1.256/−.522/.772**		**.907**	**.000**	**7.927**	**2**	**.019**
21	−1.356/−.702/−.101/.679		.907	.000	6.329	3	.097
22	−2.239/−1.443/−.653/.666		.907	.000	14.095	3	.003
23	−1.861/−1.171/−.569/.630		.907	.000	1.674	3	.643
24	−1.874/−.934/.611		.907	.000	2.213	2	.331
25	−2.254/−1.200/−.505/.663		.907	.000	2.169	3	.538
26*	−1.428/−.198/.701/2.316	−1.428/−.011/1.330/2.072	.907	.000	18.666	3	<.001

Items marked with *do not have equal thresholds across groups (i.e., a statistical significant $\Delta\chi^2$). Reference items are in bold

For item 24, three instead of four thresholds are estimated, because the fifth response category is never selected

Results

Configural invariance

Configural invariance, or invariance of the factor structure, is tested to find out whether the Belgian and Iranian respondents use similar frames of reference when completing the WHOQOL-BREF[1] [14, 19]. The fit indices are on the boundary of what is considered acceptable: $\chi^2(492) = 2087.154$, $RMSEA = .110$, and $CFI = .907$. Although these values are not unambiguously in favor of the model, we decided to proceed with the invariance tests. An alternative would be to make subtle modifications to the model (for example adding cross-loadings on the basis of the modification indices; see [11]) to increase the model fit, but this route was not pursued, because all 24 facet items are clearly designed to measure one and only one domain, which would make cross-loadings meaningless from a substantive point of view.

Factor loading invariance

In a second step, it is evaluated whether, on top of the same factor structure, factor loadings are invariant across groups. It was found that the model fit did not deteriorate considerably: $\chi^2(512) = 2088.015$ ($\Delta\chi^2(20) = 44.263$), $RMSEA = .107$, and $CFI = .907$ ($\Delta CFI = .000$). This implies that all factor loadings, except for those that were fixed to one for identification purposes (items 3, 5, 8, and 20), are invariant across both samples.

[1] The following constraints were applied in order to identify the model (see [11]): (1) the latent factor means are fixed to 0 in the first group, (2) the intercepts are fixed to 0 in all groups, (3) the scaling factors are fixed to one in the first group, (4) per factor, the factor loading of one item is fixed to 1 (these items are called the reference items), and (5) for each item, an equality constraint was set on the first threshold, and additionally for the reference items, the second threshold was also constrained across groups.

Table 3 Overview of the four domains of the WHOQOL-BREF (WHO 1998) and the cross-cultural measurement invariance of the concerned facets items

WHOQOL-BREF factor	Invariant factor loadings + invariant thresholds	Invariant factor loadings + non-invariant thresholds
Physical	**3. To what extent do you feel that physical pain prevents you from doing what you need to do?**	10. Do you have enough energy for every day life?
		15. How well are you able to get around physically?
		16. How satisfied are you with your sleep?
	4. How much do you need any medical treatment to function in your daily life?	17. How satisfied are you with your ability to perform your daily living activities?
		18. How satisfied are you with your capacity for work?
Psychological		**5. How much do you enjoy life?**
		6. To what extent do you feel your life to be meaningful?
		7. How well are you able to concentrate?
		11. Are you able to accept your bodily appearance?
		19. How satisfied are you with yourself?
		26. How often do you have negative feelings such as blue mood. despair. anxiety. depression?
Social relationships	**20. How satisfied are you with your personal relationships?**	
	21. How satisfied are you with your sex life?	
	22. How satisfied are you with the support you get from your friends?	
Environment	**8. How safe do you feel in your daily life?**	
	9. How healthy is your physical environment?	
	12. Have you enough money to meet your needs?	
	13. How available to you is the information you need in your daily life?	
	14. To what extent do you have the opportunity for leisure activities?	
	23. How satisfied are you with the conditions of your living place?	
	24. How satisfied are you with your access to health services?	
	25. How satisfied are you with your transport?	

Reference items are printed in bold

Threshold invariance

Threshold invariance is tested by constraining the factor loadings as well as the item thresholds across samples. The fit indices obtained by equating both thresholds and factor loadings were $\chi^2(579) = 2754.381$ ($\Delta\chi^2(59) = 552.285$), and $RMSEA = .118$, which indicate a badly fitting model. Moreover, the goodness-of-fit measure decreased substantially: $CFI = .874$ ($\Delta CFI = .033$).

In order to identify the non-invariant thresholds, we adopted the procedure of Cheung and Rensvold [20]. This consists of testing models where consecutively each single threshold is equated across groups, while all other thresholds are allowed to vary. In Table 2, the Δ-indices are listed, comparing the model with invariant factor loadings to each consecutive model. Since a sensitive index is needed to detect differences in thresholds of a single item, $\Delta\chi^2$ is used as a criterion. In order to account for multiple testing, a Bonferroni correction was applied, and the critical p-value is set at .002. As such, items 5, 6, 7, 10, 11, 15, 16, 17, 18, 19, and 26 are found to fail threshold invariance.

Finally, a multi-group CFA was fitted in which all thresholds were equated across samples, except for items 5, 6, 7, 10, 11, 15, 16, 17, 18, 19, and 26. This model fits better: $\chi^2(547) = 2268.245$ ($\Delta\chi^2(35) = 167.542$), $RMSEA = .108$, and CFI equals .900 ($\Delta CFI = .007$).

Overview

It is found that the items of the WHOQOL-BREF perform differently with regard to measurement invariance. All items have invariant factor loadings, while some do not

stand the test of invariant thresholds. Table 3 gives an overview of how the four domains are affected.

Discussion

The aim of this study was to evaluate to what extent the WHOQOL-BREF measures QOL similarly across nations. Our findings indicate that only 11 out of the 24 facets items have invariant factor loadings and thresholds (another four facets were fixed as reference items for identification purposes, and as a result their invariance cannot be assessed).

Hence the accuracy of cross-national comparisons with the WHOQOL-BREF must be questioned. It has been argued that one can still make meaningful comparisons even if full measurement invariance does not hold [20–26]. The logic is that if the proportion of non-invariant items is rather small, this will not heavily affect the group comparison [27]. The WHOQOL-BREF faces serious problems, since for the physical and psychological domains, there are almost no invariant items. Moreover, Cheung and Rensvold [27] extended this criterion by adding two more requirements. First, the non-invariant items should relate logically to the constructs. Second, the non-invariant items should load on the same factors for all groups (i.e., configural invariance must hold). Also, these additional requirements are a cause of concern, because support for configural invariance was far from overwhelming, as demonstrated by the moderate fit indices. As a result, it is advised that the WHOQOL-BREF be used only with due caution in cross-national comparisons. Of note, it is the physical and psychological domains of the WHOQOL-BREF that are less invariant than the social relations and environmental domains. This observation may come as a surprise, since most would probably have speculated that physical and psychological items are more "universal" than the theoretically more culturally contingent social relations and environmental items [28].

Also, our conclusions must be viewed with caution because of some limitations of this study. It concerns a comparison of Dutch-speaking Belgians (or Flemish) and Iranians, which prevents one to readily generalize to other cross-national comparisons. Second, also the use of a non-representative convenience sample limits the generalizability of the results. However, the homogeneity of the sample lends support to the hypothesis that the source of the non-invariance is national or cultural. Finally, in both countries, a different mode of administration was used. Although a number of studies have shown that administration mode is of minor importance [29–33], it cannot be ruled out that it confounded the results. Further research is needed to establish the applicability of QOL instruments for cross-national and cross-cultural comparisons.

References

1. Diener, E., & Diener, M. (1995). Cross-cultural correlates of life satisfaction and self-esteem. *Journal of Personality and Social Psychology, 68*, 653–663.
2. Diener, E., & Suh, E. M. (Eds.). (2000). *Culture and subjective well-being Cambridge*. MA: MIT Press.
3. Schimmack, U., Radhakrishnan, P., Oishi, S., Dzokoto, V., & Ahadi, S. (2002). Culture, personality, and subjective well-being: Integrating process models of life satisfaction. *Journal of Personality and Social Psychology, 82*(4), 582–593.
4. Bernheim, J. L., Theuns, P., Mazaheri, M., Hofmans, J., Flieghe, H., & Rose, M. (2006). The potential of anamnestic comparative self assessment (ACSA) to reduce bias in the measurement of subjective well-being. *Journal of Happiness Studies, 7*, 227–250.
5. Oishi, S., Diener, E. F., Lucas, R. E., & Suh, E. M. (1999). Cross-cultural variations in predictors of life satisfaction: Perspectives from needs and values. *Personality and Social Psychology Bulletin, 25*, 980–990.
6. Veenhoven, R. (1991). Is happiness relative? *Social Indicators Research, 24*, 1–34.
7. Diener, E., Suh, E. M., Smith, H., & Shao, L. (1995). National differences in reported subjective well-being: Why do they occur? *Social Indicators Research, 34*, 7–32.
8. Sam, D. L. (2001). Satisfaction with life among international students: An exploratory study'. *Social Indicators Research, 53*, 315–337.
9. Murphy, B., Herrman, H., Hawthorne, G., Pinzone, T., & Evert, H. (2000). *Australian WHOQoL instruments: User's manual and interpretation guide*. Melbourne, Australia: Australian WHOQoL Field Study Centre.
10. Skevington, S. M., Lotfy, M., & O'Connell, K. A. (2004). The World Health Organization's WHOQOL-BREF quality of life assessment: Psychometric properties and results of the international field trial. A Report from the WHOQOL Group. *Quality of Life Research, 13*, 299–310.
11. Temme, D. (2006). Assessing measurement invariance of ordinal indicators in cross-national research. In S. Diehl and R. Terlutter (Eds.), *International advertising and communication: Current insights and empirical findings* (pp. 455–472). Gabler.
12. Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39*, 479–515.
13. Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*, 466–491.
14. Hofmans, J., Pepermans, R., & Loix, E. (2009). Measurement invariance matters: A case made for the ORTOFIN. *Journal of Economic Psychology, 30*, 667–674.
15. Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: The Guilford Press.
16. Woehr, D. J., Arciniega, L. M., & Lim, D. H. (2007). Examining work ethica cross populations. A comparison of the multidimensional work ethic profile across three diverse cultures. *Educational and Psychological Measurement, 67*, 154–168.
17. Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233–255.
18. Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cut-off values for fit indexes and dangers in overgeneralizing Hu and Bentler's findings. *Structural Equation Modeling, 11*, 320–341.
19. Lievens, F., Anseel, F., Harris, M. M., & Eisenberg, J. (2007). Measurement invariance of the pay satisfaction questionnaire across three countries. *Educational and Psychological Measurement, 67*, 1042–1051.

20. Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25*, 1–27.
21. Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*, 78–90.
22. Hofmans, J., Pepermans, R., & Dries, N. (2008). The career satisfaction scale: Response bias among men and women. *Journal of Vocational Behavior, 73*, 397–403.
23. Millsap, R. E., & Kwok, O. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods, 9*, 93–115.
24. Poortinga, Y. H. (1989). Equivalence of cross-cultural data: An overview of basic issues. *International Journal of Psychology, 24*, 737–756.
25. Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456–466.
26. Yoo, B. (2002). Cross-group comparisons: A cautionary note. *Psychology & Marketing, 19*, 357–368.
27. Cheung, G. W., & Rensvold, R. B. (1998). Cross-cultural comparisons using non-invariant measurement items. *Applied Behavioral Science Review, 6*, 93–110.
28. Brown, D. E. (2000). Human universals and their implications. In N. Roughley (Ed.), *Being humans: Anthropological universality and particularity in transdisplinary perspectives* (pp. 156–174). New York: Walter de Gruyter.
29. Carini, R. M., Hayek, J. C., Kuh, G. D., Kennedy, J. M., & Ouimet, J. A. (2003). College student responses to web and paper surveys: Does mode matter? *Research in Higher Education, 44*, 1–19.
30. Pettit, F. A. (2002). A comparison of world-wide web and paper- and pencil personality questionnaires. *Behavior Research Methods, Instruments, & Computers, 34*, 50–54.
31. Pouwer, F., Snoek, F. J., van der Ploeg, M., Heine, R. J., & Brand, A. N. (1998). A comparison of the standard and the computerized versions of the well-being questionnaire (WBQ) and the diabetes treatment satisfaction questionnaire (DTSQ). *Quality of Life Research, 7*, 33–38.
32. Bushnell, M. A., Martin, M. L., & Parasuraman, B. (2003). Electronic versus paper questionnaires: A further comparison in persons with asthma. *Journal of Asthma, 40*, 751–762.
33. Cronk, B. C., & West, J. L. (2002). Personality research on the internet: A comparison of web-based and traditional instruments in take-home and in-class settings. *Behavior Research Methods, Instruments, & Computers, 34*, 177–180.