

INVESTIGATIVE REPORT

Assessing Health-related Quality of Life in Hand Eczema Patients: How to Overcome Psychometric Faults when Using the Dermatology Life Quality Index

Robert F. OFENLOCH¹, Thomas L. DIEPGEN¹, Elke WEISSHAAR¹, Peter ELSNER² and Christian J. APFELBACHER³

¹Department of Clinical and Social Medicine, University Hospital Heidelberg, Heidelberg, ²Department of Dermatology, University Hospital Jena, Jena, and ³Medical Sociology, Institute of Epidemiology and Preventive Medicine, University of Regensburg, Regensburg, Germany

Health-related quality of life (HRQOL) has become an important patient reported outcome in health service research. The dermatology life quality index (DLQI) is the most commonly used instrument in dermatology. In recent years, the psychometric properties of the DLQI have been a subject of debate as requirements of modern test theory seem not to be fulfilled. The aim of this study was to test whether those violations also occur in patients with hand eczema. We collected data of 602 hand eczema patients who participated in an inpatient dermatology rehabilitation program in Germany. In order to report meaningful scores of the DLQI, data were analysed according to the principles of modern test theory. We calibrated the DLQI using the Rasch model, resulting in a 6 item version with a range between 0–15 points. This version showed no significant misfit to the Rasch model ($p > 0.14$). By using a Rasch analysis the results were evaluated in a second sample of hand eczema patients ($n = 511$). Even if all demographic characteristic of this sample were different, we were able to replicate the results found in this study ($p > 0.21$). In conclusion, we recommend to use an alternative scoring procedure as presented in this article if the DLQI is used in hand eczema patients. Key words: HRQOL; hand eczema; DLQI; Rasch model; IRT.

Accepted Mar 5, 2014; Epub ahead of print Mar 7, 2014

Acta Derm Venereol 2014; 94: 658–662.

Robert Ofenloch, Dipl. rer. soc., Department of Clinical and Social Medicine, University Hospital Heidelberg, Thibautstr. 3, DE-69115 Heidelberg, Germany. E-mail: robert.ofenloch@med.uni-heidelberg.de

Assessing patient-reported outcomes (PROs) to evaluate the success of clinical trials has become common practice and health-related quality of life (HRQOL) has become one of the most important PROs. HRQOL is not only used as an important outcome measure in clinical trials, it can also be used for treatment evaluation or to increase patient self-awareness and empowerment (1). Additionally in the National Health Service (NHS) of the United Kingdom, HRQOL is used to determine at an individual level if a specific therapy is reimbursed for a patient. For example,

according to the National Institute of Health and Clinical Excellence (NICE), alitretinoin (9-cis-retinoic acid) is only reimbursed for therapy in hand eczema (HE) if patients have a Dermatology Life Quality Index (DLQI) score > 15 (2).

The DLQI was developed nearly 2 decades ago according to the principles of classical test theory as a skin disease-specific instrument to assess HRQOL (3). Due to its easy application and the increasing importance of HRQOL in the evaluation of clinical studies the DLQI became the most commonly used HRQOL measure in dermatology (4, 5). However, a modern paradigm of test theory called the Rasch model (RM) is now assumed to be the new standard in the development of HRQOL instruments (6). Recently, modern test theory was also applied to the DLQI in psoriasis and atopic dermatitis (AD) patients and since the DLQI failed to fulfil the strict requirements, the use of the DLQI for those diseases was criticised (7). The objective of this study was to psychometrically test the DLQI in a sample of HE patients by using the principles of modern test theory.

MATERIALS AND METHODS

Study population

In one sample of HE patients we analysed and calibrated the DLQI according to modern test theory and in a second sample of HE patients we replicated the results obtained in the first sample. Sample 1 represents patients with HE who were assessed consecutively while receiving treatment for occupational skin diseases in Germany. Data were collected at the Department of Dermatology Jena/Falkenstein and at the Department of Clinical Social Medicine Heidelberg. Sample 2 was drawn from the German Chronic Hand Eczema registry (CARPE) (8). DLQI data of the 7 largest centres were used to replicate the results obtained from subjecting sample 1 to Rasch analysis.

Determining the sample size for Rasch analysis depends on several factors and there is no tool available for power calculation. Therefore we aimed to have a sample size of 500 cases according to the guidelines of Hobart & Cano (9) which is also sufficient to analyse differential item functioning between 2 groups (10).

Item response theory – the Rasch model (see Appendix S1¹)

Statistical analysis (see Appendix S1¹)

Recalibration of the dermatology life quality index (see Appendix S1¹)

¹<http://www.medicaljournals.se/acta/content/?doi=10.2340/00015555-1842>

RESULTS

In a first step the DLQI data in sample 1 ($n=602$) was coded according to the rules reported by its developers (3) resulting in 527 eligible cases (75 cases were excluded because of at least 2 missing values according to (3)). The same was done with sample 2 ($n=565$) resulting in 511 eligible cases. The sample characteristics of both study populations are presented in Table I. All characteristics presented in Table I were significantly different between sample 1 and sample 2 ($p < 0.05$). The proportion of males and the impairments in HRQOL measured by the DLQI were higher in sample 1, while mean age and mean duration since first onset of HE were higher in sample 2.

Rasch analysis

Findings from assessing the fit of the DLQI (sample 1) to the RM are presented in Table II. Four items (items 3, 5, 7 and 8) showed significant misfit because of fit-residuals lying outside the ± 2.5 range. Seven items showed uniform differential item functioning (DIF) according to gender or age group, 2 items showed non-uniform DIF (items 7 and 10) between centres or gender and 2 items had disordered thresholds. Overall, the DLQI had a person separation index (PSI) of 0.82, indicating good reliability. However, the DLQI significantly misfitted the RM (item-trait interaction $p < 0.001$).

The distribution of the persons across the logit scale of HRQOL impairment as measured by the DLQI is presented in the upper part of Fig. 1. The lower part of Fig. 1 shows the distribution of the item thresholds on the same scale. The graphs show that the thresholds are clustered in the middle of the scale (-1.5 – 2.0) and are therefore redundant in this area. In contrast, items assessing mild impairments in HRQOL are missing (see Fig. 1 values < -1.5 , lower graph), and $> 20\%$ of the investigated HE population fall within this area (see Fig. 1 values < -1.5 , upper graph).

Recalibration of the DLQI using the Rasch model. Table III shows the frequencies of the categories (0 to 3) for the

Table I. Sample characteristics

	Sample 1 ($n=527$)	Sample 2 ($n=511$)
Gender, %		
Male	330 (62.6)	241 (43.2)
Female	197 (37.4)	316 (56.6)
Age, years, mean (SD)	44.7 (11.6)	47.3 (13.1)
Range (min–max)	49 (18–67)	60 (17–77)
Disease duration, years, mean (SD)	7.2 (7.4)	9.4 (9.6)
Range (min–max)	49 (0–49)	55 (1–56)
Dermatology Life Quality Index, mean (SD)	9.8 (6.8)	8.3 (6.1)
Range (min–max)	29 (0–29)	29 (0–29)

10 items of the DLQI in sample 1 and the corresponding thresholds in between. Items 1 and 2 showed the best balanced category frequencies and thresholds for the DLQI. For items 7 and 9 all thresholds were very close to each other and disordered, those items were therefore rescored in a first calibration step (scoring now: “not at all”=0; “a little”=1; “a lot”=1; “very much”=1; giving a scoring of “0-1-1-1” instead of “0-1-2-3”). This calibrated model still showed misfit to the RM (item-trait interaction $p < 0.001$). Therefore we calibrated the items 3 and 8 where the thresholds 2 and 3 were too close to each other (< 0.5 logits). In both cases category 3 was the smallest neighbour so we collapsed categories 2 and 3 of those items by rescoring the items. For item 6 thresholds 1 and 2 were too close. Since categories 2 and 3 were the smallest in that area the item was rescored “0-1-1-2”.

After this rescoring procedure we computed fit statistics again. The overall fit statistics for all steps of the recalibration process are presented in Table IV. The DLQI still had a good PSI of 0.83 and, although it was still misfitting the RM, fit indices had slightly improved (item-trait interaction $p > 0.001$). Individual item statistics revealed that items 5 and 8 showed significant overfit to the scale (fit residuals -3.5 and -2.9). In order to compute the next fit statistics items 5 and 8 were removed from the scale. After this adjustment the overall fit statistics again improved slightly (see Table IV). However, a look at the

Table II. Individual item statistics of the DLQI in sample 1 ($n=511$) according to the Rasch analysis

Item description	Location ^a	Fit residual ^b	χ^2	p -value ^c	Uniform DIF	Non-uniform DIF	Disordered thresholds ^d
(1) Itchy, sore, painful, or stinging	-1.40	-0.28	12.63	0.125	Gender (0.016)		
(2) Embarrassment/self-consciousness	-0.20	-1.43	6.74	0.565	Gender (0.041)		
(3) Interferes with shopping/looking after home/garden**	-0.38	-2.53	17.11	0.029	Gender (0.003)		
(4) Influences choice of clothes	0.90	-0.25	6.49	0.592			
(5) Affects social/leisure activities**	0.17	-4.33	39.59	0.000			
(6) Affects ability to do sports*	0.36	0.24	4.18	0.840	Gender (0.028)		
(7) Prevents working/studying*	-0.92	5.45	89.39	0.000		Centre (0.000)	x
(8) Creates problems with partner/close friends/relatives	0.41	-3.46	0.65	0.004	Age (0.044)		
(9) Causes sexual difficulties**	0.56	-0.75	6.55	0.585	Gender (0.000)		x
(10) Problem with treatment	0.51	-0.05	14.38	0.072	Gender (0.006)	Gender (0.026)	

*This item showed critical violations according to the Rasch model and had to be adjusted or **removed during the calibration.
^aThe lowest/highest value indicates that the corresponding item is assessing the mildest/most severe impairment. ^bFit residual have to be in a range of ± 2.5 .
^caccording to χ^2 with 7 degrees of freedom; indicating a misfit of the corresponding item if significant. ^dDisordered thresholds are indicating that the answer categories of the corresponding item are not ordinal scaled.
 DIF: differential item functioning.

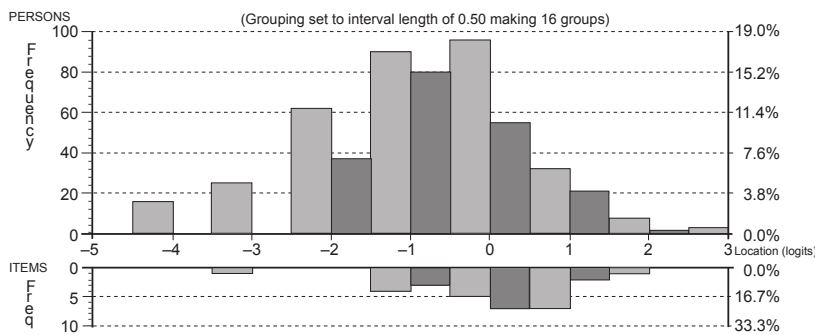


Fig. 1. Person and item threshold distribution for the DLQI. The distribution of the persons across the logit scale of HRQOL impairment is shown in the upper part. The lower part shows the distribution of the item thresholds on the same scale.

individual item statistics revealed that the fit residual for item 3 was outside of the acceptable range (-2.6). Item 9 also showed significant DIF by gender which was robust to Bonferroni correction ($p < 0.001$). Therefore items 3 and 9 were also removed from the scale.

After this adjustment the overall statistics showed that the DLQI was not significantly misfitting the RM anymore (item-trait interaction $p > 0.14$). A PSI of 0.72 indicated good internal reliability. The inspection of item statistics showed no significant overfit or overdiscrimination anymore. A marginal uniform DIF was detected for item 6 (sports) between genders indicating that males are more impaired in their sport activities than women at the same level of overall HRQOL impairment. However, this DIF was not robust to Bonferroni correction and should therefore be interpreted with caution. The alternative scoring structure for the calibrated DLQI is presented in Table V.

Evaluation of results found in sample 1

We evaluated the results obtained with sample 1 using sample 2. As in sample 1 the original scored DLQI data showed good internal reliability (PSI=0.80) but misfitted the RM significantly (item-trait-interaction < 0.001). The calibrated DLQI data of this sample did not misfit the RM significantly (item-trait interaction $p > 0.21$) and had a reasonable PSI of 0.68. All item residuals were in the ± 2.5 range and all thresholds were ordered correctly. The uniform DIF according

to gender in item 6 (sports) was significant again ($p < 0.001$) also indicating that males were more impaired in this domain compared with females. Additional uniform DIF was detected for item 7 (working) by age ($p < 0.001$) showing that the oldest age group was less impaired in this domain.

Construct validity

To test construct validity we computed a regression model with the sum scores of the original and the Rasch calibrated DLQI. The Rasch calibrated DLQI showed a strong correlation with the original DLQI ($\beta = 0.95$) and even if the Rasch calibrated DLQI consists only of 6 items with a total score range from 0–15 it explained 90% of the variance (R^2) of the original DLQI. In a next step both scores were regressed as dependent variable on physician global assessment (24), disease severity reported by the patients and gender as independent variables. As expected those variables showed only weak associations with HRQOL ($\beta = 0.16-0.29$). The largest discrepancy between the corresponding standardised regression coefficients between the 2 models was 0.03; while the R^2 for the model with the Rasch-calibrated DLQI was slightly higher (0.157 vs. 0.174). These results show that the Rasch-calibrated DLQI measures the same construct as the original DLQI.

DISCUSSION

Recently the RM was applied to the DLQI to study its psychometric properties in psoriasis and AD patients. In psoriasis patients disordered thresholds and DIF between different language versions of the DLQI were detected (25). Another study revealed disordered thresholds and DIF between psoriasis and AD patients (7). In our study similar problems in HE patients were found.

This is the first study applying the RM to the DLQI in a large sample of HE patients and replicating the findings in another large sample. Overall, misfit of the DLQI was detected as well as individual item misfit, disordered

Table III. Category frequencies and thresholds

	Score 0		Score 1		Score 2		Score 3
	Label "not at all"	Threshold 1	Label "a little"	Threshold 2	Label "a lot"	Threshold 3	Label "very much"
(1) Itchy, sore, painful, or stinging	31	-1.90	181	0.38	172	1.53	127
(2) Embarrassment/self-consciousness	161	-1.18	175	0.12	114	1.07	61
(3) Interferes with shopping/looking after home/garden	163	-1.00	161	0.29	100	0.71	87
(4) Influences choice of clothes	330	-0.62	111	-0.08	51	0.70	19
(5) Affects social/leisure activities	225	-1.00	148	0.07	89	0.92	49
(6) Affects ability to do sports	283	-0.38	113	-0.14	71	0.52	44
(7) Prevents working/studying	122	-0.13	98	0.10*	104	0.04*	187
(8) Creates problems with partner/close friends/relatives	272	-0.75	135	0.21	61	0.54	43
(9) Causes sexual difficulties	332	-0.17	98	0.23*	40	-0.06*	41
(10) Problem with treatment	263	-0.95	152	0.19	65	0.76	31

*disordered threshold. Bold indicates threshold with a distance < 0.5 logits.

Table IV. Overall fit statistics for the DLQI models during the calibration process

Step:	χ^2 item-trait interaction				Item fit residual Mean (SD)	Person fit residual Mean (SD)	Misfitting items* n
	Value	DF	p	PSI			
Primary model	202.54	70	0.000	0.82	-0.74 (2.67)	-0.27 (0.98)	4
After rescoreing 1	145.26	70	0.000	0.83	-0.77 (1.49)	-0.29 (0.91)	3
After rescoreing 2	112.13	70	0.001	0.83	-0.79 (1.71)	-0.29 (0.91)	2
After deleting item 5 and 8	77.08	56	0.032	0.79	-0.65 (1.31)	-0.27 (0.84)	2
After deleting item 3 and 9 (final model)	51.95	42	0.140	0.72	-0.60 (1.42)	-0.32 (0.82)	0
Replication of the final model	55.33	48	0.217	0.68	-0.61 (1.23)	-0.31 (0.78)	0

DF: degrees of freedom; PSI: person separation index; SD: standard deviation; *according to the individual item fit residuals.

thresholds and DIF (mainly for gender) in various items. Those results imply that DLQI scores of HE patients cannot be compared between men and women because those groups answer differently to half of the DLQI questions. Taking the disordered thresholds into account in the worst case it would be possible that a patient reporting a lower DLQI score may be more impaired in HRQOL than a patient reporting a one-point higher score. This can be very problematic since some regulatory agencies (e.g. the NICE in the UK) use the DLQI for reimbursement decisions (2).

In this study we successfully recalibrated the DLQI and made it fit the RM – overall and at item level. Though up to 2 items (one item in sample 1, 2 items in sample 2) showed uniform DIF, we have retrieved a reasonable measure since this DIF can be explained.

The possibility to explain DIF is a crucial difference in the assessment of DIF in HRQOL instruments compared to educational tests (where the RM was developed and introduced). While in educational tests (e.g. math test) a question favouring one group cannot be accepted (26) this is different in instruments assessing HRQOL. Sometimes there are items which are more relevant to one group compared to another and thereby reflecting real differences in the HRQOL impairment. We explain the DIF found in the Rasch-calibrated DLQI as follows: a) Uniform DIF for item 6 (sports) by gender: this difference can be explained by the fact that males in Germany engage more in sports compared to females at all levels of activity, which has been shown in population-based studies (27). Consequently males are more likely to be impaired in this domain even at the same level of overall HRQOL impairment. b) Uniform DIF for item 7 (working) by age:

this difference occurred only in sample 2 and is reasonable, as half of the population in the highest age group in this sample was retired already and hence less likely to be impaired in this domain – again at the same level of HRQOL impairment. This uniform DIF was not found in sample 1 since no participant in this sample was retired.

Limitations of this study

After calibrating the DLQI by removing 4 items and reducing the score from 2 of the remaining items, we were able to demonstrate construct validity for the calibrated DLQI. The regression model of the calibrated DLQI with the original scored DLQI showed nearly perfect association. Furthermore the regression models with indicators of HE severity showed nearly identical results for both DLQI versions.

According to its developers the DLQI can be analysed according to 6 headings 1: symptoms and feelings, daily activities, leisure, work and school, personal relationships and treatment. In our study all items assessing personal relationships and half of the items assessing daily activities and leisure were removed. In light of the strong correlations of the calibrated DLQI with the original DLQI and the results from the regression analysis it seems that the deleted variables have very little influence on patients’ HRQOL. This somewhat contradicts the construct validity of the original DLQI, because the assessed headings should have relevance to the HRQOL in the observed population and should not be redundant. In case of HE those results show that the deleted items are redundant and therefore not adequate to assess impairments according to the mentioned headings.

Another problem is shown in Fig. S1¹ which should be looked at in comparison with Fig. 1. The Rasch modifications have removed the redundant item thresholds in the middle of the continuum of the person and item threshold distribution. However, also the Rasch-calibrated DLQI is in need of items assessing very mild or very severe impairments – Rasch analysis can optimise psychometric properties and show deficits, but it cannot add information when items are missing.

Twiss and colleagues (7) have shown that the DLQI scores are not comparable for psoriasis and AD patients. Comparing Table II with the results from Twiss and colleagues (7) reveals that the item locations as presented here for HE patients are obviously different from those of psoriasis and

Table V. Scoring of the Rasch calibrated DLQI for hand eczema populations

	Original score and labels			
	0 “not at all”	1 “a little”	2 “a lot”	3 “very much”
(1) Itchy, sore, painful, or stinging	0	1	2	3
(2) Embarrassment/self-consciousness	0	1	2	3
(4) Influences choice of clothes	0	1	2	3
(6) Affects ability to do sports	0	1	1	2
(7) Prevents working/studying	0	1	1	1
(10) Problem with treatment	0	1	2	3

Items not reported here have been removed during the calibration process.

AD patients. This indicates that the DLQI does not measure the same (comparable scores) for these patient groups with different dermatological diseases. Hence, we recommend to use generic HRQOL instruments, such as SF36 (28) or EQ-5D (29) to compare the impact of different diseases and real disease-specific instruments as measures in clinical studies where sensitive outcome measures are needed, or as measures used for clinical recordkeeping where it is important to assess the specific HRQOL impairments patients are suffering from. The use of dermatology-specific instruments can only be recommended in skin diseases in which a disease-specific instrument for the specific dermatological disease of interest is not available. We suggest that if the DLQI is used in studies investigating HE patients the proposed alternative scoring (see Table IV or use the SPSS-syntax in Appendix S2¹) should be used to report the results.

ACKNOWLEDGEMENTS

We thank the following physicians for administering the DLQI to the patients of the 2 samples used for this study: Karl-Christian Appl (Berlin); Andrea Bauer, Jochen Schmitt (TU Dresden); Marion Büttner, Martha Fröhlich, Anna Neubauer-Tsykunova (University Hospital Heidelberg); Soo-Jin Cha (University of Jena); Anne-Katrin Dumke, Daniela Kelterer, Andrea Krauthem, Hilmar Schwantes (Clinic for Occupational Diseases Falkenstein); Ralph von Kiedrowski (Company for Medical Study & Services Selters), Vera Mahler (University of Erlangen); Sonja Molin, Thomas Ruzicka (LMU Munich).

The study was in part funded by the young scientists program of the German network 'Health Services Research Baden-Württemberg' of the Ministry of Science, Research and Arts in collaboration with the Ministry of Employment and Social Order, Family, Women and Senior Citizens, Baden-Württemberg, Germany.

The authors declare no conflict of interest.

REFERENCES

- van Cranenburgh OD, Prinsen CA, Sprangers MA, Spuls PI, de Korte J. Health-related quality-of-life assessment in dermatologic practice: relevance and application. *Dermatol Clin* 2012; 30: 323–332.
- Rodgers M, Griffin S, Paulden M, Slack R, Duffy S, Ingram JR, et al. Alitretinoin for severe chronic hand eczema: a NICE single technology appraisal. *Pharmacoeconomics* 2010; 28: 351–362.
- Finlay AY, Khan GK. Dermatology Life Quality Index (DLQI) – a simple practical measure for routine clinical use. *Clin Exp Dermatol* 1994; 19: 210–216.
- Lewis V, Finlay AY. 10 years experience of the Dermatology Life Quality Index (DLQI). *J Invest Dermatol Symp Proc* 2004; 9: 169–180.
- Basra MK, Fenech R, Gatt RM, Salek MS, Finlay AY. The Dermatology Life Quality Index 1994–2007: a comprehensive review of validation data and clinical results. *Br J Dermatol* 2008; 159: 997–1035.
- Tennant A, McKenna SP, Hagell P. Application of Rasch analysis in the development and application of quality of life instruments. *Value Health* 2004; Suppl 1: S22–S26.
- Twiss J, Meads DM, Preston EP, Crawford SR, McKenna SP. Can we rely on the Dermatology Life Quality Index as a measure of the impact of psoriasis or atopic dermatitis? *J Invest Dermatol* 2012; 132: 76–84.
- Apfelbacher CJ, Akst W, Molin S, Schmitt J, Bauer A, Weishaar E, et al. CARPE: a registry project of the German Dermatological Society (DDG) for the characterization and care of chronic hand eczema. *J Dtsch Dermatol Ges* 2011; 9: 682–688.
- Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technol Assess* 2009; 13: iii, ix–x, 1–177.
- Scott NW, Fayers PM, Aaronson NK, Bottomley A, de Graeff A, Groenvold M, et al. The practical impact of differential item functioning analyses in a health-related quality of life instrument. *Qual Life Res* 2009; 18: 1125–1130.
- Chang C-H, Reeve BB. Item response theory and its applications to patient-reported outcomes measurement. *Eval Health Prof* 2005; 28: 264–282.
- Macdonald P, Paunonen SV. A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educ Psychol Meas* 2002; 62: 921–943.
- Hambleton RK, Jones RW. Comparison of classical test theory and item response theory and their applications to test development. *Educational measurement: issues and practice* 1993; 12: 38–47.
- Reise SP, Waller NG. Item response theory and clinical measurement. *Annu Rev Clin Psychol* 2009; 5: 27–48.
- Nilsson AL, Tennant A. Past and present issues in Rasch analysis: The functional independence measure (FIMTM) revisited. *J Rehabil Med* 2011; 43: 884–892.
- Masters GN. A Rasch model for partial credit scoring. *Psychometrika* 1982; 47: 149–174.
- Marais I, Andrich D. Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *J Appl Meas* 2008; 9: 200–215.
- Andrich D. An expanded derivation of the threshold structure of the polytomous Rasch model that dispels any “threshold disorder controversy”. *Educ Psychol Meas* 2012; 73: 78–124.
- Zumbo BD. *A handbook on the theory and methods of differential item functioning (DIF)*. Ottawa: National Defense Headquarters, 1999.
- Tennant A, Penta M, Tesio L, Grimby G, Thonnard JL, Slade A, et al. Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: the PRO-ESOR project. *Medical Care* 2004; 42: I.
- Bond TG, Fox CM. *Applying the Rasch model: Fundamental measurement in the human sciences*: Lawrence Erlbaum; 2001.
- Linacre JM. Investigating rating scale category utility. *J Outcome Meas* 1999; 3: 103–122.
- Pallant JF, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *Br J Clin Psychol* 2007; 46: 1–18.
- Coenraads PJ, Van Der Walle H, Thestrup-Pedersen K, Ruzicka T, Dreno B, De La Loge C, et al. Construction and validation of a photographic guide for assessing severity of chronic hand dermatitis. *Br J Dermatol* 2005; 152: 296–301.
- Nijsten T, Meads DM, de Korte J, Sampogna F, Gelfand JM, Ongenaes K, et al. Cross-cultural inequivalence of dermatology-specific health-related quality of life instruments in psoriasis patients. *J Invest Dermatol* 2007; 127: 2315–2322.
- Zwick R. A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *Research report ETS [RR-12-08] 2012 [cited 2013]; Available from: <http://www.ets.org/Media/Research/pdf/RR-12-08.pdf>*.
- Rütten A, Abu-Omar K, Lampert T, Ziese T. Körperliche Aktivität. *Gesundheitsberichterstattung des Bundes* 2005; 26.
- Ware JE, Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care* 1992; 30: 473–483.
- Brooks R. EuroQol: the current state of play. *Health Policy* 1996; 37: 53–72.