

Comparing the impact of psoriasis and atopic dermatitis on quality of life: co-calibration of the PSORIQoL and QoLIAD

J. Twiss · S. P. McKenna

Accepted: 16 January 2014
© Springer International Publishing Switzerland 2014

Abstract

Background Disease-specific patient-reported outcome (PRO) measures are designed to be highly relevant to one disease. It is widely believed that comparisons of outcomes between patients with different diseases are only possible using generic measures. The present study employs a novel method of using Rasch analysis to co-calibrate scores from different disease-specific PRO measures, allowing scores to be compared across diseases.

Methods Psoriasis patients ($n = 146$, mean age = 44.4, males = 50 %) completed the Psoriasis Quality of Life scale (PSORIQoL) and atopic dermatitis patients ($n = 146$, mean age = 45.5, males = 50 %) the Quality of Life in Atopic Dermatitis scale (QoLIAD). Both measures employ the needs-based model of QoL, and they share five common items—providing a link between assessments. The groups were analysed separately, and then combined to test fit to the Rasch model.

Results Both scales showed good fit to the Rasch model after minor adjustments (PSORIQoL: $\chi^2 p = 0.25$; QoLIAD: $\chi^2 p = 0.51$). For the combined dataset, one common item showing differential item functioning by disease was removed and fit to the Rasch model was achieved ($\chi^2 p = 0.08$). The co-calibrated scale successfully distinguished between perceived severity groups ($p < 0.001$).

Conclusions It is possible to co-calibrate scores on the PSORIQoL and QoLIAD. This is one of the first studies in health research to demonstrate how Rasch analysis can be used to make comparisons across diseases using different disease-specific measures. Such an approach maintains the

greater relevance and, consequently, accuracy associated with disease-specific measurement.

Keywords Psoriasis · Atopic dermatitis · PSORIQoL · QoLIAD · Rasch analysis · Co-calibration · Patient-reported outcomes · Quality of life

Introduction

When selecting a patient-reported outcome (PRO) measure for a specific disease, researchers are presented with an array of choices [1, 2]. The type of PROs available can be split broadly into three categories: Generic measures such as the Medical Outcomes Study Short-Form 36 General Health Survey (SF-36) [3] and the Nottingham Health Profile (NHP) [4]; speciality generic measures used across (for example) several dermatology conditions such as the Dermatology Life Quality Index (DLQI) [5] and Skindex [6] and disease-specific measures such as the Psoriasis Quality of life scale (PSORIQoL) [7] and Quality of Life in Atopic Dermatitis scale (QoLIAD) [8]. In dermatology, a large amount of research has been conducted recently evaluating available scales [9–13].

Generic PRO measures, as the name implies, are intended for use with any disease. They usually provide a profile of scores covering different outcome constructs. Such measures are often favoured in research studies for practical reasons; as they are familiar and readily available for use—negating the need to develop a new scale. Traditionally, it is also believed that generic PRO measures can be used when pooling data from different diseases, when comparing scores between diseases or when comparing data with normative values [14, 15]. Patients with different skin diseases are frequently combined in research

J. Twiss (✉) · S. P. McKenna
Galen Research Ltd, Enterprise House, Manchester Science
Park, Lloyd Street North, Manchester M15 6SE, UK
e-mail: jtwiss@galen-research.com

studies in this way [16–21]. Generic PRO measures, such as the DLQI and SF-36, are used in these studies.

Recent evidence has raised serious doubts about the validity of using generic PRO measures to combine patients with different diseases as items in the scales have often been found to exhibit differential item functioning (DIF) by disease. This has been observed for generic PRO measures such as the SF-36 [22, 23] and for speciality generic PRO measures such as the DLQI [12]. Coupled with this is evidence that many popular generic PRO measures used in dermatology have weak measurement properties [9, 10, 12, 24].

Disease-specific measures offer many potential benefits over generic outcomes. They are designed to assess the issues that are important to patients with a specific disease so they are more relevant and acceptable [25]. Evidence also shows that disease-specific PRO measures are more responsive than generic outcome measures in a range of conditions [26–29]. A criticism of disease-specific PRO measures is that they cannot be used to combine or compare data from patients with different diseases. However, through the application of item response theory (IRT), it is now possible to do this.

The use of IRT may make it possible to administer relevant disease-specific PRO measures to patients with different conditions and then co-calibrate the specific measures onto the same measurement continuum [30]. Such an approach is commonly used in educational settings to equate tests of different ability levels, but no published example could be found in health research that uses this methodology to link outcomes from different diseases. In order to allow this form of analysis, it is necessary for the disease-specific outcome measures to be based on the same theoretical model and to have overlapping items.

This study aimed to co-calibrate scores from two disease-specific outcome measures; the Psoriasis Quality of Life scale (PSORIQoL) and the Quality of Life in Atopic Dermatitis scale (QoLIAD) so that scores from the measures could be combined and compared validly. Both of these scales employ the needs-based model of QoL [31].

Methods

Participants

Secondary analyses were conducted on data collected from atopic dermatitis (AD) and psoriasis patients in the UK. The psoriasis sample was recruited from the Manchester Psoriasis Service database at Salford Royal Hospital, Greater Manchester. The AD sample was recruited through the National Eczema Society ($n = 286$). A random sample

of 73 males and 73 females was selected from the latter sample to match the number of males and females in the psoriasis sample. The samples were matched as uneven samples may affect the robustness of analysis of variance (ANOVA) assessments (used during the differential item functioning assessments).

Measures

The PSORIQoL [7] and QoLIAD [8] are measures of QoL specific to psoriasis and atopic dermatitis, respectively. Both employ the needs-based model of QoL [31]. The model is built on theoretical work on human needs [32–34]. According to the model, QoL is high when needs are fulfilled and low when they are not. Conceptualising QoL in this way clearly sets the construct apart from impairments and activity limitations. These latter types of outcome may influence QoL but are only important to the patient insofar as they prevent need fulfilment.

Relevant needs may fall into several different categories including those related to safety and security, socialisation, affection, esteem, cognition and personal development. Although individual illnesses will influence different needs, there is commonly overlap. For example, patients with quite different diseases such as rheumatoid arthritis and multiple sclerosis may both have difficulty participating with family and friends—depriving them of the need for social interaction. In this context, the impact of different diseases on need fulfilment will have both common and unique elements. The aim of the present study was to test whether needs common to two different diseases could be used as anchors to co-calibrate the unique aspects onto the same metric.

Both scales were developed using the Rasch model [35], a simple logistic latent trait IRT model, to ensure unidimensionality and good measurement properties. Each measure consists of 25 items with dichotomous response options (True/Not true). High scores on the measures indicate poor QoL.

Procedure

Three stages were involved in the analysis:

1. PSORIQoL and QoLIAD data were analysed separately to test fit to the Rasch model to re-establish their scaling properties.
2. Secondly, data from both surveys were combined using a common item design (described below) and Rasch analysis fit statistics for the common items and overall scale were investigated.
3. Finally, the validity of the co-calibrated scale was then investigated by relating Rasch scores from the co-calibrated scale to disease type and perceived disease severity.

Rasch analysis

The Rasch model was selected due to its special mathematical properties, in particular, its satisfaction of fundamental measurement requirements [36]. According to the model, the probability that an individual will respond in a certain way to a particular item is a logistic function of the relative distance between the item-location parameter and the person-location parameter. This difference governs the probability of the expected response for a person, of a given ability (level of QoL), to an item of a given severity. If the observed data do not deviate significantly from the expected responses, then the items fit the Rasch model. Several different kinds of fit statistic are used to assess fit to the Rasch model.

Analyses conducted were consistent with published guidelines [37]. Internal reliability was analysed using the Person Separation Index (PSI). The PSI is indicative of the power of the items to distinguish between respondents. A PSI score of 0.70 is the minimum acceptable level. The overall scale fit to the model was examined by reference to the overall item–trait interaction χ^2 fit value. A significant χ^2 statistic ($p < 0.05$) indicates that there is a real deviation of the scale from the expected pattern and a lack of fit to the Rasch model. Overall fit of the data is also investigated via item and person interaction statistics. These assessments measure the extent to which observed item and person estimates deviate from the expected. The mean location of the items is always anchored at 0. Within this function, both person fit residual and item fit residual statistics are transformed by RUMM to approximate Z-scores; representing a standardised normal distribution. When the data fit the model, the overall distribution statistics for item fit and person fit should have a mean of approximately 0 and a standard deviation of approximately 1.

Individual item fit statistics were also investigated via χ^2 fit statistics. A significant χ^2 fit statistic ($p < 0.05$ (Bonferroni adjusted)) indicates model misfit. Individual item-fit residuals were also investigated. These should fall within ± 2.5 if all individuals respond in the anticipated way. High negative residuals are indicative of item redundancy and high positive residuals multidimensionality, both types of misfit.

A requirement of the Rasch model is that items should be invariant across groups. This is investigated through tests of DIF. Differential item functioning represents instability in the order of severity of items and indicates that the scale may not work in the same way in different sub-groups of individuals (such as those defined by age or gender) that share the same level of trait being measured [38]. An ANOVA of standardised residuals was carried out to examine DIF by gender (males vs. females) and age group (below the median age vs. median age and above).

For the combined analyses, DIF by disease (AD vs. psoriasis) was investigated for the common items to ensure that these anchor items worked in the same way in both disease samples. An ANOVA p value of < 0.05 (Bonferroni adjusted) was considered to be indicative of DIF.

Targeting of items to the QoL of respondents was assessed by examining person–item distribution graphs. These show the ordering of both persons and items on the same logit scale and indicate whether the items in the scale are well matched to the respondents.

If a scale is unidimensional, it should be free from item dependency; similarity in items such that answers to one item have too strong an influence over answers to other items [37]. This is assessed by inspecting the residual correlation matrix for pairs of items with correlations exceeding 0.3.

The unidimensionality of the measure is further assessed with an additional test [39]. If a scale is unidimensional, the responses to any subset of items within a scale should give the same estimate of person ability, given appropriate targeting. Responses to two subsets of answers (loading most differently on the first factor of the Residual Principal Components Analysis) were assessed to see whether they produced significantly different estimates for each participant (using t tests). Twelve items were selected (providing 12 response thresholds) that loaded positively on the first factor of the Residual Principal Components Analysis and compared with 12 items that loaded negatively on the first factor. If fewer than 5 % of these tests are statistically significant, the measure may be considered to be unidimensional. A 95 % binomial confidence interval is provided.

A single indicator of item misfit or DIF does not necessarily warrant the removal of an item. During the analyses if an item showed minor misfit from the model, the issue was explored in more detail. The study of Rasch model fit is a process where no individual fit statistic is either necessary or sufficient [40, 41]. Therefore, the final interpretation of fit is a collective one based on several indices. Clear misfit is reported in the results section.

The Rasch Unidimensional Measurement Model (RUMM) 2030 [41] programme was employed for the analyses.

Co-calibration of the QoL scales

The scales have five items in common and so were completed by all respondents. The remaining items in each scale were completed by the relevant patients only. The responses to the common or ‘anchor items’ provide a link for the two scales. Using Rasch analysis, it is possible to estimate ability/item parameters despite missing responses and thus to co-calibrate two scales where only a proportion of the items have been completed by both samples.

Table 1 Sample characteristics

	Psoriasis (<i>n</i> = 146)	Atopic dermatitis (<i>n</i> = 146)
<i>Gender (%)</i>		
Male	73 (50.0)	73 (50.0)
Female	73 (50.0)	73 (50.0)
<i>Age (years)</i>		
Mean (SD)	44.4 (14.7)	45.5 (16.6)
Range	66.0 (17–83)	62.0 (20–82)
<i>Disease duration (years)</i>		
Mean (SD)	20.9 (13.5)	28.2 (17.5)
Range	67.0 (2–69)	76.0 (0–76)
<i>Marital status (%)</i>		
Married/living as	89 (61.0)	95 (65.1)
Single	57 (39.0)	50 (34.2)
<i>Perceived general health (%)</i>		
Excellent	16 (11.0)	15 (10.3)
Good	68 (46.6)	96 (65.8)
Fair	44 (30.1)	32 (21.9)
Poor	17 (11.6)	3 (2.1)
<i>Perceived illness severity (%)</i>		
Mild	38 (26.0)	60 (41.1)
Moderate	57 (39.0)	57 (39.0)
Quite severe	43 (29.5)	26 (17.8)
Very severe	6 (4.1)	3 (2.1)

Validity of co-calibrated scale

Further analyses were conducted using the Rasch-transformed scores to assess the validity of the co-calibrated scale. Scores for psoriasis and atopic dermatitis patients were compared (using a *t* test). In addition, scores for psoriasis and atopic dermatitis patients were combined and compared by perceived severity group (mild, moderate, quite severe and very severe) using ANOVA. *p* values of <0.01 were used to indicate statistical significance.

Rasch logit scores for the co-calibrated scale were linearly transformed to a 0–100 scale.

Results

Demographic details of the samples are shown in Table 1.

Rasch analysis of individual scales

Table 2 reports overall fit statistics for the PSORIQoL and QoLIAD. Both measures fit ($\chi^2 p > 0.05$) the Rasch model initially. The PSIs were good indicating that the measures had appropriate levels of reliability. The additional test of unidimensionality for the QoLIAD suggested marginal

multidimensionality. This was investigated further by assessing items with high fit residual correlations (indicating local dependency). Three items showed evidence of local dependency. Removing two of these improved unidimensional statistics (Table 2).

Table 3 includes the fit statistics for the individual items in the QoLIAD and PSORIQoL. No clear item misfit was identified.

Items in both scales were well matched to the severity level of the patients. Minimal levels of item redundancy were observed. The PSORIQoL items assessing greatest level of trait (the most severe QoL items) were ‘*I don’t value myself*’ and ‘*I just want to shut myself away.*’ The item capturing the lowest level of the trait (the mildest item) was ‘*I constantly have to look after my skin.*’ The logit coverage (the distance between most severe and mildest items) was 6.2. The most severe QoLIAD item was ‘*I can’t concentrate on anything else*’ and the mildest item ‘*I worry about my appearance.*’ The total logit coverage was 6.6.

Rasch analysis of combined samples

As only minor inconsistencies were detected in the individual scales, all items were retained for the combined analysis. Item–person interaction and item–trait interaction for the combined scales are shown in Table 2. Overall fit to the model was achieved (item–trait interaction $\chi^2 > 0.05$). None of the individual items exhibited misfit to the model.

The logit positions of the common linking items are shown in Table 4. DIF analyses of the linking items showed that one exhibited DIF by disease ($p = 0.00098$ —‘*I worry that people will not accept me*’ (Fig. 1). AD participants were more likely to affirm the item at each level of the trait. This item was removed from the analysis as it is necessary for linking items to be free from disease bias. Overall fit to the model changed little after the removal of the item (Table 2).

Targeting of the items to the patients is displayed in Fig. 2. The graph shows that the items had a good spread and were well matched to the person locations on the logit scale. Acceptably low end effects were observed as follows: 23 participants (7.9 %) scored the minimum and 10 participants (3.4 %) the maximum possible on the scale. The graph also indicates minimal item redundancy in the scale. The logit coverage of the combined sample was 8.3.

Validity of co-calibrated scale

The PSI value for the co-calibrated scale (0.89) was good showing that the scale was reliable (Table 2). Figure 3 shows mean scores and corresponding 95 % confidence intervals based on the standard error of the mean for disease and perceived severity groups. Comparison of samples

Table 2 Overall fit statistics

	Item–trait interaction χ^2 p value	No of items	PSI	Item–person fit residuals				Unidimensionality (confidence interval)
				Item fit residual		Person fit residual		
				Mean	SD	Mean	SD	
<i>QoLIAD</i>								
Initial fit	0.28	25	0.91	−0.25	1.12	−0.16	0.54	0.089 (0.054–0.124)
Removal of 2 items	0.51	23	0.82	−0.25	1.12	−0.15	0.53	0.048 (0.013–0.083)
<i>PSORIQoL</i>								
Initial fit	0.25	25	0.94	−0.22	1.12	−0.36	0.82	0.048 (0.013–0.083)
<i>Combined sample</i>								
Initial fit (All items included)	0.09	45	0.88	−0.19	1.15	−0.26	0.69	–
Item showing DIF by disease removed	0.08	44	0.89	−0.18	1.17	−0.26	0.68	–
Ideal values	>0.05	–	>0.85	0.00	1.00	0.00	1.00	<0.05 (Lower CI < 0.05)

Table 3 Individual item fit statistics for the separate PSORIQoL and QoLIAD analyses

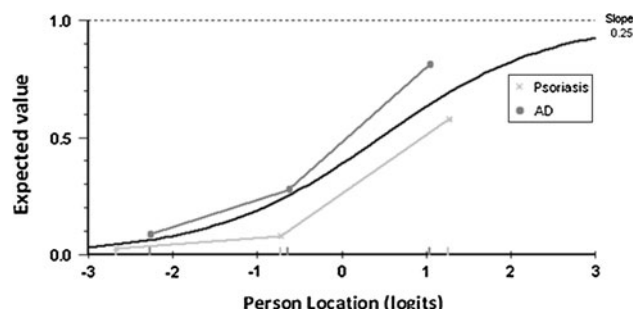
PSORIQoL					QoLIAD				
Item	Location	Fit residual	χ^2	p value	Item	Location	Fit residual	χ^2	p value
1	1.36	−1.29	1.21	0.55	1	0.97	−0.54	0.26	0.88
2	2.67	−0.87	0.74	0.69	2	2.19	−1.16	2.52	0.28
3	1.26	−0.59	0.72	0.70	3	−0.42	−0.67	3.11	0.21
4	−0.95	−0.73	2.68	0.26	4	−0.91	−2.17	5.65	0.06
5	2.08	−0.74	0.81	0.67	5	1.85	−0.58	0.93	0.63
6	−1.35	−1.44	4.49	0.11	6	−2.38	−0.55	3.58	0.17
7	0.64	−0.54	0.07	0.97	7	0.82	−0.35	0.80	0.67
8	−1.75	0.60	2.83	0.24	8	−0.51	−0.57	2.99	0.22
9	1.29	0.00	1.04	0.59	9	0.49	−0.18	0.69	0.71
10	0.12	−1.52	2.80	0.25	10	1.12	−0.43	0.64	0.73
11	−0.65	−0.10	0.26	0.88	11	−0.84	1.90	9.01	0.01
12	2.62	0.39	2.06	0.36	12	1.01	−0.42	0.27	0.87
13	−3.34	1.43	3.11	0.21	13	−0.43	−0.64	0.58	0.75
14	0.90	0.01	0.37	0.83	14	−1.89	−0.33	1.64	0.44
15	1.03	−1.45	2.47	0.29	15	−0.86	1.47	6.22	0.04
16	−0.44	−0.51	3.30	0.19	16	−0.64	3.15	1.01	0.60
17	−1.68	2.84	2.31	0.32	17	−0.24	−1.37	2.10	0.35
18	0.66	0.51	3.01	0.22	18	−1.12	0.70	3.54	0.17
19	1.45	−0.52	0.58	0.75	19	4.32	−0.46	0.63	0.73
20	1.63	−1.60	2.82	0.24	20	−0.26	−0.77	1.77	0.41
21	−2.13	1.43	5.43	0.07	21	−1.11	0.68	0.97	0.61
22	−3.46	1.51	5.59	0.06	22	0.01	−1.52	4.70	0.10
23	−0.54	−0.46	0.52	0.77	23	−0.73	−0.18	0.16	0.92
24	−1.46	−1.36	5.70	0.06	24	−0.96	−0.57	0.53	0.77
25	0.01	−0.55	1.57	0.46	25	0.51	−0.85	1.07	0.59

showed that psoriasis patients had significantly higher scores than those with atopic dermatitis ($t = 4.54$, $p < 0.001$). For the combined psoriasis and atopic

dermatitis sample, the co-calibrated scale was able to distinguish significantly between disease-severity groups ($F = 22.44$, $p < 0.001$).

Table 4 Logit positions of the common linking items

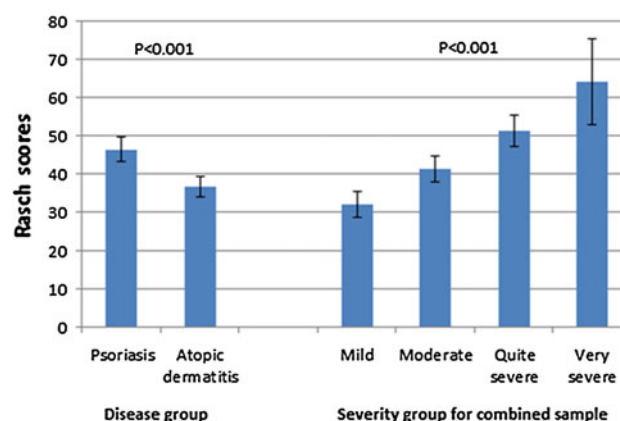
Item description	Logit position
I just want to shut myself away	2.51
I cannot bear anyone touching me	2.03
I have no self confidence	1.23
I am embarrassed about my appearance	-0.85

**Fig. 1** DIF by disease results for item 'I worry that other people will not accept me' for the combined sample

Discussion

This is the first published study showing how Rasch analysis can be used to co-calibrate scores on two different disease-specific PROs. Other studies have used Rasch analysis and other IRT models to link different scales within the same disease area [42–44]. The present study takes this approach a stage further by linking across diseases.

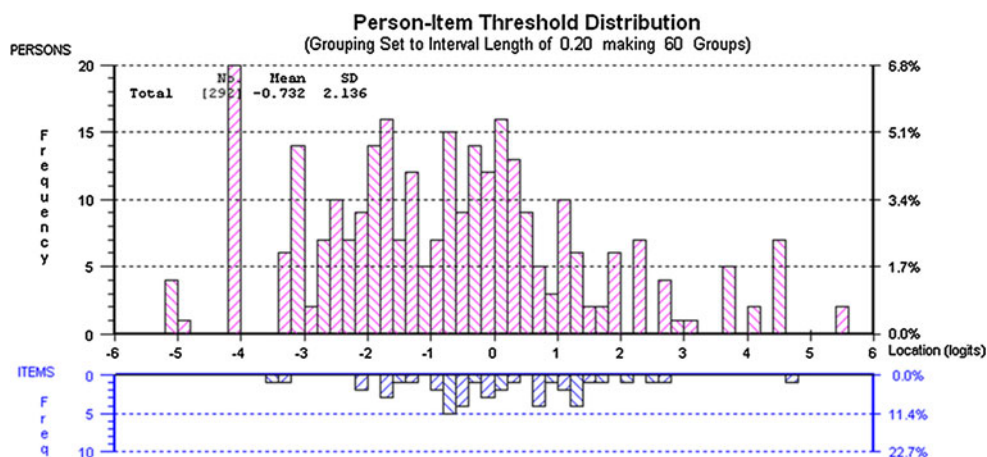
Only minor adjustments were needed to the two scales to achieve fit to the Rasch model. The final co-calibrated scale showed good measurement properties over a range of analyses. The validity of the co-calibrated scale was established by relating scores to perceived severity data

**Fig. 3** Mean Rasch scores by disease and perceived severity

collected from both samples. When data from the two samples were combined, the co-calibrated scale was able to distinguish between perceived disease-severity groups. This shows that the co-calibrated scale is valid and indicates that the approach will have value in future research studies including both psoriasis and atopic dermatitis patients.

Co-calibration of disease-specific measures has succinct advantages over relying on generic PRO measures. Disease-specific measures provide more detailed, relevant and responsive information. Due to the nature of generic measures, they are likely to miss issues of importance in specific diseases. For example, the SF-36 does not contain any items relating to self-perception, which has been identified as an important issue for the QoL of psoriasis patients [7]. In addition, many generic measures have weak measurement properties and exhibit DIF by disease [9, 10, 12, 22–24].

The new approach could be valuable in research studies that are designed to look at the impact of a specific intervention on samples consisting of patients with different diseases [16–21]. Another potential application of this

Fig. 2 Targeting of items for the combined sample

method is in comparative effectiveness studies that currently employ generic outcome measures. Such studies are used to make decisions about the allocation of scarce health resources [45]. Essentially, it provides the opportunity to compare the relative effectiveness of treatment interventions in different conditions on a common QoL metric. A head-to-head comparison of the psychometric properties of generic versus co-calibrated disease-specific measures would be beneficial in future studies.

Disease-specific measures can only be co-calibrated if they meet strict criteria. The scales must be unidimensional, based on the same measurement model and have overlapping items that are free from DIF by disease. This precludes the use of many disease-specific measures as they are often developed with unique or poorly defined conceptual frameworks. However, scales that employ the needs-based model of QoL meet these criteria as they are clearly operationalised, cover a wide variety of diseases and have good measurement properties [46–49]. Over twenty disease-specific outcome measures have been developed using the needs-based approach offering scope for co-calibration across several different diseases. The proportion of common items between the measures varies, with up to ten common items available. Co-calibration of needs-based measures for several rheumatic conditions is underway currently. It is possible that these could also be co-calibrated with the dermatological measures used in the present study. However, it remains to be tested whether this method can be applied across different disease groups where the impact on needs could be quite different. Diseases that are very different in nature have less overlap making co-calibration more challenging. In addition, DIF by disease is a potential problem as the same item may be understood and valued differently in different conditions.

The Rasch model was selected in the present study as data that fit the Rasch model meet the requirements of fundamental measurement [36]. Scales that do not fit the Rasch model cannot be co-calibrated in the way described in this study using Rasch analysis. Other, less restrictive, item response theory models are also available for co-calibrating measures [50–52].

The results showed that patients in the psoriasis sample had significantly higher scores on the co-calibrated scale than the atopic dermatitis patients indicating that they had worse QoL on average. This finding is supported by the perceived severity data that showed that a greater proportion of AD patients reported their illness severity as mild. The observed difference in severity is likely to reflect the different recruitment routes. Psoriasis patients were recruited from a local hospital, while the AD patients were members of a self-help organisation.

Limitations of the study should be noted. The samples were drawn from different sources. Despite this, the co-

calibration procedure was successful. Further investigation using different samples would be useful for assessing whether these results are replicable.

Four items were available for co-calibration of the two measures. No agreement exists on the number of items needed to perform such co-calibrations. Previous researchers have suggested that the more items available to act as anchors the more robust the co-calibration exercise will be [53, 54]. Some researchers have argued that 20 % of the test size is optimal [55]. However, others have shown that successful co-calibration can be achieved with relatively few items if the common and unique items are of good enough quality to ensure good estimates of ability [56]. The four items used in the present study were from different locations along the logit scale and exhibited good fit statistics. In addition, all other fit statistics supported the success of the co-calibration.

This study showed it was possible to link the PSORIQoL and QoLIAD measures using Rasch analysis. Only minor changes were required to allow the measures to be co-calibrated. This approach has several practical applications in research studies where patients with different diseases are combined or need to be compared. This method allows relevant, disease-specific information to be collected while also allowing scores to be compared across the two diseases. The findings provide the basis for future work linking needs-based scales across a range of different diseases.

Acknowledgments The study was unfunded. We thank the patients who completed the PSORIQoL and QoLIAD, Professor CEM Griffiths from Salford Royal Hospital and the UK National Eczema Society.

References

- Otuki, M. F., Reis, R. C., Cabrini, D., Prudente, A. S., Hori-nouchi, C. D., & Correr, C. J. (2011). Patient-reported outcomes in psoriasis research and practice. *British Journal of Dermatology*, 165(6), 1361–1362.
- Rehal, B., & Armstrong, A. W. (2011). Health outcome measures in atopic dermatitis: a systematic review of trends in disease severity and quality-of-life instruments 1985–2010. *PLoS One* 13, 6(4), e17520.
- Ware, J. E., & Sherbourne, C. D. (1992). The MOS 36-Item Short-Form Health Survey (SF-36). *Medical Care*, 30, 474–483.
- Hunt, S. M., McEwan, J., & McKenna, S. P. (1985). Measuring health status: a new tool for clinicians and epidemiologists. *Journal of the Royal College of General Practitioners*, 35(273), 185–188.
- Finlay, A. Y., & Khan, G. K. (1994). Dermatology Life Quality Index (DLQI)-a simple practical measure for routine clinical use. *Clinical and Experimental Dermatology*, 19, 210–216.
- Chren, M. M., Lasek, R. J., Flocke, S. A., & Zyzanski, S. J. (1997). Improved discriminative and evaluative capability of a refined version of Skindex, a quality of-life instrument for patients with skin diseases. *Archives of Dermatology*, 133, 1433–1440.

7. McKenna, S. P., Cook, S. A., Whalley, D., Doward, L. C., Richards, H. L., Griffiths, C. E., et al. (2003). Development of the PSORIQoL, a psoriasis-specific measure of quality of life designed for use in clinical practice and trials. *British Journal of Dermatology*, 149, 323–331.
8. Whalley, D., McKenna, S. P., Dewar, A. L., Erdman, R. A., Kohlmann, T., Niero, M., et al. (2004). A new instrument for assessing Quality of Life in Atopic Dermatitis (QoLIAD). *British Journal of Dermatology*, 150, 274–283.
9. Nijsten, T. E., Sampogna, F., Chren, M. M., & Abeni, D. D. (2006). Testing and reducing skindex-29 using Rasch analysis: Skindex-17. *Journal of Investigative Dermatology*, 126, 1244–1250.
10. Nijsten, T. E., Meads, D. M., de Korte, J., Sampogna, F., Gelfand, J. M., Ongenaes, K., et al. (2007). Cross-cultural inequivalence of dermatology-specific health related quality of life instruments in psoriasis patients. *Journal of Investigative Dermatology*, 127, 2315–2322.
11. Both, H., Essink-Bot, M. L., Busschbach, J., & Nijsten, T. (2007). Critical review of generic and dermatology-specific health-related quality of life instruments. *Journal of Investigative Dermatology*, 127, 2726–2739.
12. Twiss, J., Meads, D. M., Preston, E. P., Crawford, S. R., & McKenna, S. P. (2012). Can we rely on the Dermatology Life Quality Index (DLQI) as a measure of the impact of psoriasis or atopic dermatitis? *Journal of Investigative Dermatology*, 132(1), 76–84.
13. Prinsen, C. A., de Korte, J., Augustin, M., Sampogna, F., Salek, S. S., & Basra, M. K., et al. (2013). Measurement of health-related quality of life in dermatological research and practice: outcome of the EADV Taskforce on Quality of Life. *Journal of the European Academy of Dermatology and Venereology*, 27(10), 1195–203.
14. Chamlin, S. L., & Chren, M. M. (2010). Quality-of-life outcomes and measurement in childhood atopic dermatitis. *Immunology and Allergy Clinics of North America*, 30(3), 281–288.
15. McKenna, S. P. (2011). Measuring patient-reported outcomes: Moving beyond misplaced common sense to hard science. *BMC Medicine*, 9, 86.
16. Potocka, A., Turczyn-Jabłońska, K., & Kieć-Swierczyńska, M. (2008). Self-image and quality of life of dermatology patients. *International Journal of Occupational Medicine and Environmental Health*, 21(4), 309–317.
17. Potocka, A., Turczyn-Jabłońska, K., & Merecz, D. (2009). Psychological correlates of quality of life in dermatology patients: The role of mental health and self-acceptance. *Acta Dermatovenereologica Alpina, Pannonica et Adriatica*, 18(2), 53–58.
18. Ludwig, M. W., Oliveira Mda, S., Muller, M. C., & Moraes, J. F. (2009). Quality of life and site of the lesion in dermatological patients. *Anais Brasileiros de Dermatologia*, 84(2), 143–150.
19. Quandt, S. A., Schulz, M. R., Vallejos, Q. M., Feldman, S. R., Verma, A., Fleischer, A. B., et al. (2008). The association of dermatologist-diagnosed and self-reported skin diseases with skin-related quality of life in Latino migrant farmworkers. *International Journal of Dermatology*, 47(3), 236–241.
20. Papoutsaki, M., Chimenti, M. S., Costanzo, A., Talamonti, M., Zangrilli, A., Giunta, A., et al. (2007). Adalimumab for severe psoriasis and psoriatic arthritis: An open-label study in 30 patients previously treated with other biologics. *Journal of the American Academy of Dermatology*, 57(2), 269–275.
21. Schmitt, J., Heese, E., Wozel, G., & Meurer, M. (2007). Effectiveness of inpatient treatment on quality of life and clinical disease severity in atopic dermatitis and psoriasis vulgaris: A prospective study. *Dermatology*, 214(1), 68–76.
22. Dallmeijer, A. J., de Groot, V., Roorda, L. D., Schepers, V. P., Lindeman, E., van den Berg, L. H., et al. (2007). Cross-diagnostic validity of the SF-36 physical functioning scale in patients with stroke, multiple sclerosis and amyotrophic lateral sclerosis: A study using Rasch analysis. *Journal of Rehabilitation Medicine*, 39(2), 163–169.
23. Taylor, W. J., & McPherson, K. M. (2007). Using Rasch analysis to compare the psychometric properties of the short form 36 physical function score and the health assessment questionnaire disability index in patients with psoriatic arthritis and rheumatoid arthritis. *Arthritis & Rheumatism (Arthritis Care & Research)*, 57(5), 723–729.
24. Jenkinson, C., Fitzpatrick, R., Garratt, A., Peto, V., & Stewart-Brown, S. (2001). Can item response theory reduce patient burden when measuring health status in neurological disorders? Results from Rasch analysis of the SF-36 physical functioning scale (PF-10). *Journal of Neurology, Neurosurgery and Psychiatry*, 71, 220–224.
25. Velanovich, V. (1998). Comparison of generic (SF-36) vs. disease-specific (GERD-HRQL) quality-of-life scales for gastroesophageal reflux disease. *Journal of Gastrointestinal Surgery*, 2(2), 141–145.
26. Shiklar, R., William, M. K., Okun, M. M., Thompson, C. S., & Revicki, D. A. (2006). The validity and responsiveness of three quality of life measures in the assessment of psoriasis patients: results of a phase II study. *Health and Quality of Life Outcomes*, 27(4), 71.
27. Angst, F., Verra, M. L., Lehmann, S., & Aeschlimann, A. (2008). Responsiveness of five condition-specific and generic outcome assessment instruments for chronic pain. *BMC Medical Research Methodology*, 25(8), 26.
28. Dawson, J., Boller, I., Doll, H., Lavis, G., Sharp, R., Cooke, P., et al. (2012). Responsiveness of the Manchester-Oxford Foot Questionnaire (MOXFQ) compared with AOFAS, SF-36 and EQ-5D assessments following foot or ankle surgery. *The Journal of Bone and Joint Surgery*, 94(2), 215–221.
29. Jull, A., Parag, V., Walker, N., & Rodgers, A. (2010). Responsiveness of generic and disease-specific health-related quality of life instruments to venous ulcer healing. *Wound Repair and Regeneration*, 18(1), 26–30.
30. Reeve, B. B., Hays, R. D., Chang, C., & Perfetto, E. M. (2007). Applying item response theory to enhance health outcomes assessment. *Quality of Life Research*, 16, 1–3.
31. Hunt, S. M., & McKenna, S. P. (1992). The QLDS: A scale for the measurement of quality of life in depression. *Health Policy*, 22, 307–319.
32. Maslow, A. H. (1970). *Motivation and personality* (2nd ed.). New York: Harper & Row.
33. Max-Neef, M. A., Elizalde, A., & Hopenhayn, M. (1991). *Human scale development: conception application and further reflections* (p. 18). New York: Apex Press.
34. Kenrick, D. T., Griskevicius, V., Neuberg, S. L., & Schaller, M. (2010). Renovating the pyramid of needs: Contemporary extensions built upon ancient foundations. *Perspectives on psychological science*, 5, 292.
35. Rasch, G. (1960). (Reprinted 1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: University of Chicago Press.
36. Tennant, A., McKenna, S. P., & Hagell, P. (2004). Application of Rasch analysis in the development and application of quality of life instruments. *Value in Health*, 7(Suppl 1), S22–S26.
37. Tennant, A., & Gonaghan, P. G. (2007). The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis and Rheumatism*, 57, 1358–1362.
38. Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. New Jersey: Lawrence Erlbaum Associates.
39. Smith, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal

- component analysis of residuals. *Journal of Applied Measurement*, 3(2), 205–231.
40. Andrich, D. (1988). *Rasch models for measurement*. Sage University paper series on Quality Application in the social sciences Sage Publications Inc: Newbury Park.
 41. Andrich, D., Sheridan, B., & Luo, G. (2009). *Interpreting RUMM2030* (4TH ed.). Perth: RUMM Laboratory Pty Ltd.
 42. Latimer, S., et al. (2012). Co-calibration of deliberate self harm (DSH) behaviours: towards a common measurement metric. *Psychiatry Research*, 200(1), 26–34.
 43. Thissen, D., Varni, J. W., Stucky, B. D., Liu, Y., Irwin, D. E., & DeWalt, D. A. (2011). Using the PedsQL™ 3.0 Asthma Module to obtain scores comparable with those of the PROMIS Pediatric Asthma Impact Scale (PAIS). *Quality of Life Research*, 20, 1497–1505.
 44. Crane, P. K., Narasimhalu, K., Gibbons, L. E., Mungas, D. M., Haneuse, S., Larson, E. B., et al. (2008). Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. *Journal of Clinical Epidemiology*, 61(10), 1018–1027.
 45. Chalkidou, K. & Anderson, G. (2009). Comparative effectiveness research: international experiences and implications for the United States. [http://www.academyhealth.org/files/publications/CER_International_Experience_09%20\(3\).pdf](http://www.academyhealth.org/files/publications/CER_International_Experience_09%20(3).pdf) September, 2013.
 46. McKenna, S. P., Meads, D. M., Doward, L. C., Twiss, J., Pokrzywinski, R., Revicki, D., et al. (2011). Development and validation of the living with chronic obstructive pulmonary disease questionnaire. *Quality of Life Research*, 20, 1043–1052.
 47. Doward, L. C., McKenna, S. P., Meads, D. M., Twiss, J., & Eckert, B. J. (2009). The development of patient-reported outcome indices for multiple sclerosis (PRIMUS). *Multiple Sclerosis*, 15, 1092–1102.
 48. Keenan, A. M., McKenna, S. P., Doward, L. C., Conaghan, P. G., Emery, P., & Tennant, A. (2008). Development and validation of a needs-based quality of life instrument for osteoarthritis. *Arthritis & Rheumatism (Arthritis Care & Research)*, 59(6), 841–848.
 49. McKenna, S. P., Doughty, N., Meads, D. M., Doward, L. C., & Pepke-Zaba, J. (2006). The Cambridge pulmonary hypertension outcome review (CAMPHOR): A measure of health-related quality of life and quality of life for patients with pulmonary hypertension. *Quality of Life Research*, 15(103–115), 54.
 50. Haley, S. M., Ni, P., Lai, J. S., Tian, F., Coster, W. J., Jette, A. M., et al. (2011). Linking the activity measure for post acute care and the quality of life outcomes in neurological disorders. *Archives of Physical Medicine and Rehabilitation*, 92, S37–S43.
 51. McHorney, C. A. (2002). Use of item response theory to link 3 modules of functional status items from the asset and health dynamics among the oldest old study. *Archives of Physical Medicine and Rehabilitation*, 83(3), 383–394.
 52. Thissen, D., Varni, J. W., Stucky, B. D., Liu, Y., Irwin, D. E., & DeWalt, D. A. (2011). Using the PedsQL 3.0 asthma module to obtain scores comparable with those of the PROMIS pediatric asthma impact scale (PAIS). *Quality of Life Research*, 20(9), 1497–1505.
 53. Vale, C. D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement*, 10, 333–344.
 54. Wolfe, E. W. (2000). Equating and item banking with the Rasch model. *Journal of Applied Measurement*, 1(4), 409–434.
 55. Angoff, W.H. Scales, norming, and equivalent scores. In R. L. Thorndike (Ed.), *Education measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.
 56. Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement*, 8, 347–364.