



Equivalence of measurement properties of the Dermatology Life Quality Index (DLQI) across 13 European countries

Sam Salek¹ · Jeffrey Johns² · Faraz M. Ali² · Florence J. Dalgard³ · Jörg Kupfer⁴ · Andrew Y. Finlay²

Received: 11 July 2025 / Revised: 30 July 2026 / Accepted: 29 August 2026
© The Author(s) 2026

Abstract

The Dermatology Life Quality Index (DLQI) is the most widely used tool to measure burden of skin diseases and assess intervention effectiveness. It is embedded in national guidelines and disease registries in >45 countries and available in >138 translations. This study examined psychometric properties of the DLQI across 13 European languages. 3,408 patients had completed the DLQI questionnaire with no missing data. Confirmatory factor analysis (CFA) was performed using datasets of each country and multigroup CFA (MGCFA) across all countries. Known-group analysis was performed of DLQI sum-score against physician-assessed severity by country. All countries' DLQI data showed essential unidimensionality with good fits to a 1-factor CFA model with CFI, TLI unscaled values all >0.97 (COSMIN criteria). MGCFA indicated different country translations of the DLQI conceptualize the same construct and showed equality of factor loadings. Although, scalar invariance was not seen, meaning that direct comparison of mean DLQI scores across 12 countries is not supported by the current results, partial scalar invariance was achieved after constraining one threshold for DLQI item 7. Known-group analysis of DLQI versus disease severity was significant for every country (all $p < 0.025$). Good psychometric properties were present across all 13 languages examined with essential unidimensionality confirmed. The DLQI differentiated between levels of disease severity and EQ-5D-3 L levels across all these languages. This adds further confidence in the appropriateness of using the DLQI with this range of languages.

Keywords Dermatology Life Quality Index · DLQI · Measurement equivalence · Invariance · Multigroup CFA · MGCFA

Introduction

The Dermatology Life Quality Index (DLQI) [1] is the most widely used dermatological quality of life measure globally. It is available in more than 138 translations [2], has been

incorporated in guidelines or registries in >45 countries [3] and has been used in 457 randomised controlled trials [4] (as primary outcomes in >24 [5]). A systematic review reported on the DLQI used as benchmark for validating >100 other newly developed patient reported outcome measures [6]. The extensive use of the DLQI does not, by itself, guarantee optimal measurement in all populations and disease severities. Our recently published item-level diagnostics for this complete dataset which showed that no specific items contribute disproportionately to non-invariance [7]. However, further validation and investigation in various populations, settings and diseases is needed.

Reported limitations of the DLQI include floor effects, dimensionality and cultural differential item functioning DIF, as was highlighted in our previous systematic review [8]. These are being addressed in a series of studies we are conducting and publishing. However, that systematic review also revealed good test–retest reliability, internal consistency reliability, responsiveness to change, known

✉ Jeffrey Johns
johnsj4@cardiff.ac.uk
Sam Salek
m.s.salek@herts.ac.uk

¹ School of Health, Medicine and Life Sciences, University of Hertfordshire, Hatfield AL10 9AB, UK

² Division of Infection and Immunity, School of Medicine, Cardiff University, Cardiff CF14 4XN, UK

³ National Center for Dual Diagnosis, Innlandet Hospital Trust, Brumundal N-2381, Norway

⁴ Institute of Medical Psychology, Justus Liebig University, Friedrichstraße 36, 35392 Gießen, Germany

group validity, correlation with other patient reported outcomes (PRO)/quality of life (QoL) measures and interpretability and clinical meaningfulness of the scores.

COSMIN embeds unidimensionality in the definition of structural validity i.e. “the degree to which the scores of an outcome measurement instrument are an adequate reflection of the dimensionality of the construct to be measured”, and not as an independent metric [9]. COSMIN interprets that a scale is unidimensional if statistical analysis shows that the items reflect one underlying construct (e.g. by CFA, EFA, IRT or Rasch). Structural validity is the measurement property used to evaluate whether a set of items is unidimensional and provides evidence that all items in a scale measure a single latent construct.

For measures such as the DLQI, it is important to establish unidimensionality, which means that one single common factor accounts for all item covariances i.e. the measure has only one construct or dimension (latent trait) and thus all the items measure the same concept. If items instead measure multiple unrelated constructs, then summing them produces a statistically unstable and conceptually ambiguous score that does not represent any single outcome. Psychometric theory i.e. classical test theory, factor analysis, and item response theory all converge on the principle that a scale that is intended to yield a single interpretable number must demonstrate at least “essential unidimensionality” (dominant dimension with minor secondary factors) [10].

In practice, very few QoL/PRO measures are perfectly unidimensional. Health outcomes are inherently multifactorial (e.g., symptoms, functioning, emotions), so PROMs often contain clusters of related but distinct content. Modern psychometric approaches (e.g., bifactor models, Rasch residual analysis, multidimensional IRT) acknowledge that some secondary dimensions are expected in real data. However, what matters is whether a dominant general factor is strong enough that the composite score behaves like a measure of one construct. If a single score is to be calculated from all the items, the fundamental requirement of any patient-reported outcome measure is that an instrument shows essential unidimensionality, not necessarily perfect unidimensionality. This essential unidimensionality (dominant dimension with minor secondary factors), which is the standard recommended by contemporary measurement science i.e. a single score is valid only if the instrument shows essential unidimensionality, not necessarily perfect unidimensionality.

Measurement equivalence refers to the broader goal that a measure assesses the same underlying construct in the same way across different groups, cultures, times, or conditions. For example, it asks whether a questionnaire has the same meaning and interpretation across countries. For example, it asks whether a questionnaire has the same meaning and

interpretation across countries. Measurement invariance is the specific statistical framework, usually tested using confirmatory factor analysis (CFA) or structural equation modeling (SEM), to evaluate whether measurement equivalence exists. Invariance testing proceeds hierarchically: configural invariance examines whether the same factor structure exists across groups; metric (weak) invariance tests whether factor loadings are equal, allowing comparison of associations between constructs; scalar (strong) invariance tests whether item intercepts (or thresholds for categorical items) are equal, permitting comparison of latent means; and strict invariance additionally requires equal residual variances. Together, these levels provide evidence that observed group differences reflect true differences in the underlying construct rather than measurement artefacts. It should be noted that measurement equivalence more broadly encompasses issues such as translation quality, cultural relevance, conceptual equivalence, and response styles, which may not be fully captured by CFA-based invariance testing alone.

It is important to determine measurement invariance across translations if comparisons are to be made between different groups using these language variants. Measurement invariance will tell us whether individuals in different groups interpret items the same way, whether item responses have the same psychometric meaning across groups and whether latent factors are comparable (means/variances) [11].

All “approved” DLQI translations including those used in this study dataset, followed standard methodology i.e. language versions had full linguistic validation, with a minimum of two parallel independent forward translations, then co-operative agreement, then two independent back translations assessed by the Cardiff team. There was then a process of further forward and backward translations until the researchers were satisfied that the translation was as accurate as possible. Continuity over detailed assessment was ensured for all translations being conducted and approved by only three Cardiff team members over the last 30 years. The cross-cultural adaptation process described underpins semantic, experiential and conceptual equivalence. In addition, certificates are generated on approval of translations, and robustness of translation has been established for many translations by the length of use, over decades, of each translation. Additionally, when the original version was created from patient generated qualitative data, attention was paid to ensure simplicity, readability, clarity and parsimony (low complexity, brevity of 10 items) [12, 13]. This has contributed to ease of cross-cultural adaptation to other languages/settings, maintaining robustness as demonstrated by extensive multinational usage [3, 4, 8, 14].

This study analysed the data taken from a European study [15] to demonstrate unidimensionality of the DLQI within

each cross-cultural adaptation (translation) and to investigate unidimensionality and measurement invariance i.e. that the DLQI construct is measuring in the same way across the different countries. The DLQI translations used in this study have been officially recognised by the DLQI authors and were cross-culturally adaptive using accepted forward/backward translation and cognitive debriefing methods. However, few of these have undergone psychometric validity testing. Validation of quality of life questionnaires is critical to ensure they accurately and reliably measure what they intend to measure [16].

Methods

This study analysed data taken from a European multicentre cross-sectional study [15] of patients recruited from dermatological out-patient clinics in 13 countries (Belgium, Denmark, France, Germany, Hungary, Italy, Netherlands, Norway, Poland, Russia, Spain, Turkey and the UK). Two countries (Norway and Italy) had two data collection centres each and their data have been combined for the analyses. In each dermatology outpatient clinic, 250 consecutive adult out-patients were recruited. The dataset contains DLQI raw scores, EQ-5D-3 L, EQ VAS (visual analogue scale) and physician assessed disease severity. Each country administered the DLQI and EQ-5D-3 L in their own validated local language version. Patients with missing DLQI responses were excluded from analysis.

For both scoring the DLQI and for exploratory and confirmatory modelling, “Not relevant” responses (possible for 8 of the 10 items) were scored according to the official/original DLQI scoring system. This states that a response of “not relevant” is scored as 0, the same as for “not at all” [1]. In a clinic consultation, however, interpretation of “not relevant” and “not at all” would require a different conversation with the patient.

As Pearson correlation has a reliance on the assumption of continuous and normally distributed data that often renders it inadequate for ordinal data [17], and the DLQI items are ordinal (e.g. Likert scales), inter-item correlation was analysed with polychoric correlations [18]. The interpretation is slightly different from Pearson correlations, where inter-item polychoric correlation values are ideally between 0.20 and 0.65 indicating good correlation, with no values < 0.20 (indicating weak relationship) while extremely high inter-item correlations (e.g., > 0.85) might suggest multicollinearity or item redundancy [19, 20]. For convergent validity, Spearman’s correlations between DLQI scores and EQ VAS, age and physician assessed severity, and among these variables, with data pooled across countries. Pooled data increases statistical power, provides generalisability,

captures greater variability and gives the reader greater confidence in the analysis conclusions.

Confirmatory factor analysis (CFA) was performed using R scripts for version 4.2.2 (R Foundation for Statistical Computing) library Lavaan and JASP version 0.95.4 to examine whether the data fit the predetermined 1-factor model [21]. Because the DLQI data were not normally distributed and showed floor-effects, the Weighted Least Squares with Mean and Variance adjusted (WLSMV) estimator was used with the mean and variance adjustment procedure [22] and with the delta parametrization [23]. An analysis of polychoric correlations among latent responses [24] was also used, that is designed for ordinal data, is asymptotically distribution free, and does not make distributional assumptions about the observed variables [19, 25]. Fit parameters for CFA are described in the supplementary data.

Reliability (Cronbach’s alpha [26, 27], Revelle’s omega hierarchical [28], Bentler’s omega total [29, 30] and McDonald’s omega total [31]) were calculated with semTools in R. Internal consistency was determined by Cronbach’s alpha with item deletion [26, 27] calculated with IBM SPSS Statistics version 27, and correlations and effect sizes were all calculated in R.

Model invariance or measurement invariance refers to the idea that psychological constructs like quality of life (QoL) are measured in the same way across different groups e.g. countries. It ensures that comparisons between groups are meaningful—that differences in scores reflect actual differences in the construct, not biases in how the items behave. Model invariance can be investigated by CFA using data from several groups simultaneously with multiple-group confirmatory factor analysis (MGCFA), or exact measurement invariance approach. MGCFA runs a single model and all the global fit statistics are estimated based on the data from all the groups simultaneously. A key feature is its ability to constrain parameters across groups and test if they are equal.

Different levels of measurement invariance can be determined where each level adds constraints and tests whether items behave similarly across the groups. These levels are hierarchical with higher levels imposing more restrictions on the measurement parameters, but at the same time allow a higher degree of comparability. At configural level, the constraints tested pertain to different groups having the same factor structure, giving interpretation of whether the groups conceptualize the construct similarly (with residual variances fixed to 1.0, so all thresholds can be freely estimated across groups). At metric level, equality of factor loadings is investigated i.e. are the items equally related to the latent factor across groups. At scalar level, for ordinal data equal thresholds are investigated to determine if group differences in scores are meaningful i.e. whether mean scores across 12

countries studied can be compared [32]. MGCFA computes differences between model fit indices using the chi-square difference test and/or invariance. MGCFA was performed in R with the lavaan, semTools, and semPlot packages. As there were no responses for categories 3 and 4 for item DLQI10 in the France dataset and the data set was small relative to the others ($N=107$), the France dataset was not included in the MGCFA models and thus only 12 countries were included. If the condition of configural invariance is fulfilled, one adds cross-group equality constraints on the factor loadings (metric invariance) and subsequently on the intercepts (scalar invariance) [33, 34]. As metric invariance was established, but full scalar invariance was not supported, partial scalar invariance was examined. Non-invariant item thresholds were identified through inspection of large modification indices (or score statistics), small p -values, and threshold/loading parameter estimates. Equality constraints on the identified non-invariant thresholds were then relaxed sequential, allowing these parameters to vary across groups while retaining invariance constraints on the remaining thresholds.

Adequate model fit for structural validity was defined according to COSMIN criteria [35] as a Comparative Fit Index (CFI) or Tucker-Lewis Index (TLI) >0.95 , a Root Mean Square Error of Approximation (RMSEA) <0.06 , and/or a Standardized Root Mean Square Residual (SRMR) <0.08 . Internal consistency was assessed using Cronbach's alpha, with values ≥ 0.70 considered acceptable. With large sample sizes, the Chi-square test can become very sensitive, and even small differences between the model and the data can result in a significant p -value ($p < 0.05$). In such cases, it is often better to rely on other fit indices, such as the Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMR). The Scaled Comparative Fit Index (Scaled CFI) is an alternative version of the Comparative Fit Index (CFI), designed to provide a more accurate model fit assessment when using robust estimation methods in structural equation modelling (SEM), particularly when data are non-normal or when using the MLR (Maximum Likelihood with Robust Standard Errors) estimator in lavaan. When data are non-normally distributed (i.e., they have skewed distributions or outliers), the standard CFI may be biased or inaccurate. The Scaled CFI corrects for these issues by scaling the fit index, making it more robust in the presence of non-normality. Scaled CFI adjusts the robust CFI further to account for the non-central distribution of the Chi-square statistic when using robust estimation methods. It scales the robust CFI value to correct for the non-centrality. The conflation of WLSMV estimation compared to MLR-based robustness is discussed in the supplementary data.

Known group analysis was performed of the DLQI sum score against physician assessed severity with data pooled across countries. The Jonckheere-Terpstra non-parametric statistical test [36, 37] in SPSS Statistics version 27 (an extension of the Mann-Whitney U test, similar to Kruskal-Wallis, but specifically testing for a priori ordered alternatives) was used to detect ordered trends (i.e. groups follow a monotonic trend across the independent groups). The assumptions of the dependent variable being measured at the ordinal level, the independent variable consisting of two or more categories, independent groups, independence of observations (no relationship between the observations in each group or between the groups themselves), and the distribution of scores for each group of the independent variable having the same shape, were all upheld.

Results

The number of patients included in the study and distribution of mean DLQI, EQ VAS and physician assessed severity scores, and age across the 13 European countries is shown in Table 1. The mean DLQI scores varied from 3.46 (Belgium) to 10.69 (Poland), mean EQ VAS from 47.8 (Hungary) to 76.4 (Spain), and mean disease severity from 1.38 (Russia and Belgium) to 1.89 (Poland). Spearman's correlations between DLQI and other measures were EQ VAS -0.409 , disease severity 0.407 , EQ-5D-3 L mobility 0.211 , EQ-5D-3 L selfcare 0.257 , EQ-5D-3 L activity 0.370 , EQ-5D-3 L pain 0.409 , EQ-5D-3 L anxiety and depression 0.365 (all $p < 0.001$) indicating that these variables were suitable as anchors.

Distributions of DLQI, EQ VAS and age are different across countries (Independent-Samples Kruskal-Wallis Tests, all $p < 0.001$). The distribution of physician assessed severity is different across countries (Chi-square test, $p < 0.001$).

Gender distribution was consistent across countries with all data showing slightly more females than males (except for Germany) (Figure S1). Overall, 55.8% of patients in the dataset were female. Disease severity reported by physicians was mild 33.8%, moderate 42.2% or severe 14.8%, with 11.1% missing. No severity data was reported for the Netherlands. Individual distributions across countries are shown in Figure S2.

Overall, psoriasis was the most commonly diagnosed disease within countries, comprising 10.0% (Turkey) to 27.2% (Russia) of total patients. There were also high levels of specific diseases reported in some countries: acne (30.0% Turkey), venous insufficiency/leg and other ulcers (29.9% Hungary), non-melanoma skin cancer (25.5% Netherlands, 21.4% Denmark, 19.7% Norway), and hidradenitis

Table 1 Distribution of mean DLQI, EQ VAS and physician assessed severity scores, and age across the 13 European countries

Country	N	Mean DLQI (range)	Mean EQ VAS (range)	Mean age	Mean severity (range)
Poland	233	10.69 (0–30)	62.86 (0–100)	43.99	1.89 (1–3)
Russia	248	10.51 (0–30)	63.44 (0–100)	48.76	1.38 (1–3)
Italy	498	8.16 (0–29)	69.61(10–100)	43.58	1.85 (1–3)
Hungary	204	7.37 (0–30)	47.80 (2–100)	48.24	1.87 (1–3)
Germany	243	7.01(0–30)	64.36 (0–100)	52.19	1.44 (1–3)
Spain	249	6.84 (0–20)	76.44 (12–100)	44.97	1.54 (1–3)
Norway	469	6.70 (0–30)	69.28 (0–100)	49.56	1.86 (1–3)
Turkey	250	6.02 (0–25)	73.57 (20–100)	30.70	1.68 (1–3)
Denmark	233	5.75 (0–30)	67.94 (30–100)	52.85	1.79 (1–3)
Netherlands	203	5.12 (0–25)	56.14 (8–100)	52.18	-
UK	253	5.03 (0–29)	75.35 (10–100)	45.90	1.66 (1–3)
France	107	4.52 (0–26)	63.91 (30–100)	49.02	1.53 (1–3)
Belgium	218	3.46 (0–19)	69.59 (25–100)	48.40	1.38 (1–3)
Total	3408				

Table 2 Correlations of variables averaged by country

Spearman's correlations	R	Test statistic	p-value	95% CI
DLQI and EQ VAS	− 0.409	8.1E+09	<0.0001	[− 0.438, − 0.379]
DLQI and age	− 0.119	7.0E+09	<0.0001	[− 0.154, − 0.089]
DLQI and severity	0.292	3.3E+07	<0.0001	[0.256, 0.325]
EQ VAS and age	− 0.201	6.6E+09	<0.0001	[− 0.236, − 0.168]
EQ VAS and severity	− 0.185	6.8E+07	<0.0001	[− 0.222, − 0.150]
Age and severity	0.023	5.3E+07	0.235	[− 0.015, 0.056]

suppurativa (11.2% Denmark). A full disease distribution plot by country is shown in Figure S3.

All DLQI polychoric inter-item correlations within each country data were above 0.42 (values <0.20 indicate weak relationship), and most were <0.65 (average 0.55) indicating moderate to strong associations. No extremely high correlations (>0.85) were observed that might indicate multicollinearity or item redundancy [19, 20].

Spearman's correlations between DLQI scores and EQ VAS, age and physician assessed severity, and among these variables, are shown in Table 2. Correlations were significant for all except age and severity with R values >0.2. The correlations were moderate for DLQI/EQ VAS, and weak for DLQI/severity and EQ VAS/age. Physician assessed severity may therefore be suitable as an anchor for further analysis of DLQI scores. However, correlations were very weak for EQ VAS/age and EQ VAS/severity, and age and severity were not correlated ($R=0.023$, $p=0.235$).

There was good Spearman's correlation for all countries between the DLQI scores and physician assessed severity (Table 3). All correlation coefficients are significant ($p<0.05$) with most having $p<0.001$. This variable is therefore appropriate to be used as an anchor in further analysis. The results of known-group analysis of DLQI score vs. physician assessed severity are also shown in Table 3. Known group validity analysis was significant ($p<0.05$) between groups for all countries with most having significance at $p<0.001$.

When DLQI scores were converted to DLQI Score Meaning Banding [38], by country, a wide range of score distribution patterns was observed (Fig. 1). However, DLQI scores showed a general inverse exponential distribution from low to high DLQI scores for all countries (Figure S4).

Unidimensionality of the DLQI has previously been shown in this dataset [7]. The results of CFA using 1-factor model by country is shown in Table 4. Acceptability criteria of at least good fit was met by all country data, but some country-specific indices suggest local misfit, including RMSEA>0.10 for France and elevated SRMR for France, Turkey and Belgium. For the France dataset there were no responses for categories 3 and 4 and the data set was small ($N=107$). SRMR is unique among global fit indices because it is a residual-based index and it directly summarizes the standardized difference between observed and model-implied correlations. This makes SRMR highly sensitive to country-specific anomalies, even when the overall model fits well. SRMR will inflate if an item's thresholds differ substantially from model expectations, a loading behaves differently, an item has unusual category endorsement patterns or cross-group response styles differ (e.g. extreme responding, avoidance of certain options) [39, 40]. In our datasets for France, Turkey and Belgium the SRMR elevation was due to the skewed data and sparse response categories. This instability leads to large residuals, which SRMR captures, but CFI/RMSEA do not penalize this strongly and it is not large enough to influence CFI/TLI/RMSEA globally.

A test of measurement invariance was conducted using multiple-group confirmatory factor analysis (MGCFA). Chi-Squared Difference Tests for all model comparisons were significant ($p<0.001$), indicating that each added

Table 3 Correlation of DLQI scores with physician assessed severity, known group comparisons and effect sizes between severity levels

Country	N	Correlation		Known-group DLQI vs. severity		Level comparisons	DLQI vs. severity effect size		
		^a Spearman rho	Spearman <i>p</i> -value	^b Std J-T statistic	<i>p</i> -value		^c Cohen's d	CI Lower	CI Upper
Belgium	173	0.205	0.007	2.67	0.008	Mild to moderate	0.25	− 0.09	0.58
						Moderate to severe	0.58	0.14	1.02
Denmark	223	0.332	<0.001	4.99	<0.001	Mild to moderate	0.69	0.38	1.00
						Moderate to severe	0.27	−0.07	0.62
France	91	0.268	0.010	7.58	0.01	Mild to moderate	0.06	− 0.41	0.53
						Moderate to severe	0.89	0.29	1.49
Germany	210	0.314	<0.001	4.63	<0.001	Mild to moderate	0.50	0.21	0.79
						Moderate to severe	0.78	0.27	1.29
Hungary	191	0.163	<0.001	2.29	0.022	Mild to moderate	0.49	0.14	0.84
						Moderate to severe	0.08	− 0.26	0.43
Italy	497	0.426	<0.001	9.63	<0.001	Mild to moderate	0.59	0.39	0.79
						Moderate to severe	0.72	0.47	0.97
Norway	445	0.190	<0.001	4.03	<0.001	Mild to moderate	0.43	0.21	0.66
						Moderate to severe	0.10	− 0.13	0.34
Poland	233	0.350	<0.001	5.45	<0.001	Mild to moderate	0.65	0.34	0.96
						Moderate to severe	0.59	0.20	0.98
Russia	214	0.241	<0.001	3.51	<0.001	Mild to moderate	0.42	0.13	0.71
						Moderate to severe	0.13	− 0.33	0.60
Spain	249	0.142	0.03	2.25	0.025	Mild to moderate	0.41	0.14	0.69
						Moderate to severe	− 0.07	− 0.52	0.38
Turkey	250	0.310	<0.001	4.98	<0.001	Mild to moderate	0.67	0.40	0.94
						Moderate to severe	0.16	− 0.28	0.60
UK	253	0.400	<0.001	6.421	<0.001	Mild to moderate	0.78	0.51	1.06
						Moderate to severe	0.37	− 0.06	0.79
Total	3029	0.292	<0.001	16.26	<0.001	Mild to moderate	0.90	0.82	0.98
						Moderate to severe	1.11	1.00	1.22

^aSpearman correlations, comparison made with Jonckheere-Terpstra statistic, and Cohen's-d effect sizes calculated with SPSS 29.0

Correlations calculated by Spearman rank-order correlation as a measure of the degree of monotonic association between the ordinal grouping variable (ordered groups), and the ordinal dependent variable i.e. ordinal by ordinal

An estimate of the variability (or standard error) of the correlation coefficient (for large samples) was calculated not assuming the null hypothesis i.e. calculated under the alternative hypothesis, not assuming that the true correlation is zero

^bFor large N, Jonckheere-Terpstra is approximated by a normal distribution

^cCohen's-d effect sizes are between mild to moderate, and moderate to severe levels on alternate columns. Effect sizes are as follows: 0.2–0.5=small effect (italic), >0.5=moderate effect (bold), >0.8=large effect (bolditalic)

Data for the Netherlands not calculated as physician assessed severity data was not collected

constraint significantly worsens model fit. However, Chi-square is very sensitive to sample size, so changes in CFI and RMSEA were investigated. The global fit indices are presented in Table 5.

Configural, metric (loadings) scalar (thresholds) and strict (means) invariance models demonstrated good fits to data ($\chi^2 p < 0.001$, CFI and TLI scaled > 0.95, 0.92 for means. Although the scaled RMSEA seems somewhat high (0.089) for the baseline model fit with the WLSMV estimator, a ULS estimator (which does not use a chi square statistic), gives a value of 0.061. The WLSMV uses polychoric correlations, thresholds for each item category and a diagonally weighted least squares chi-square statistic with corrections. This generates a larger χ^2 , particularly when categories are sparse (such as for the four category scores for the DLQI

items), groups differ in threshold structure, sample sizes are moderate and items show response heterogeneity or DIF. It is well known that RMSEA behaves poorly under WLSMV, especially in multi group settings, as WLSMV must estimate group-specific thresholds, thereby increasing model complexity: when degrees of freedom increase sharply, χ^2 goes up even faster and RMSEA becomes artificially inflated [38, 39, 10, 44]. Thus an RMSEA of 0.088 under WLSMV is neither unusual nor disqualifying. RMSEA is over-sensitive to minor residual misfit under WLSMV with categorical data and small residuals in polychoric matrices, small differences in thresholds across groups and slight cross-group item behaviour differences, all resulting in disproportionately large χ^2 under WLSMV.

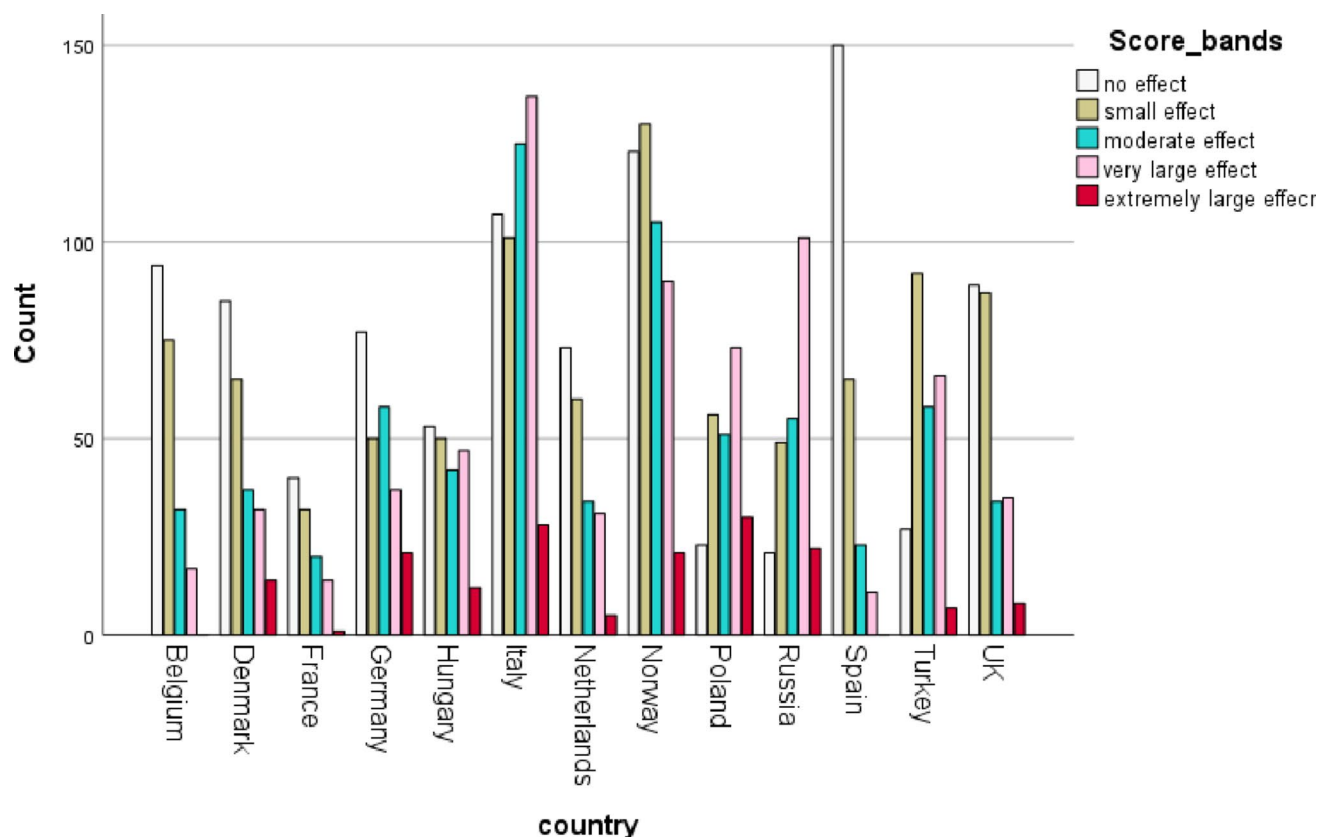


Fig. 1 DLQI score meaning banding by country (Hongbo et al. [38])

No increase in CFI and a decrease of 0.009 in RMSEA between configural and metric invariance models is within the range recommended by Chen [44], therefore, the more constrained model showed comparable fit and given reasons of parsimony, the metric invariance model has fewer estimated parameters and therefore should be accepted. In contrast, the scalar invariance model was rejected, because the decrease in CFI of -0.03 and increase in RMSEA of 0.023 were greater than the thresholds of 0.01 and 0.015 respectively. By the same criteria, mean invariance was also rejected, as expected, as the data show clear differences in means between countries due to different disease and severity of disease distributions (Figures S2 and S3).

A non-invariant item threshold for DLQI item 7 was identified (modification indices Chi-squared=39.7), small p-value (<0.0001), and by threshold/loading parameter estimates. This can be clearly seen from Figure S7. Plot of thresholds by country, where the t1 threshold for Russia (-0.53) and Turkey (-0.93) are clearly outliers in the model fit. On constraining this threshold, partial scalar variance was achieved (Table 5). Constraining all thresholds for item 7 did not improve model fit compared to constraining only threshold t1.

Discussion

There is growing interest in the use and reporting of PROMs in routine clinical practice [14] as well as in clinical research. Therefore, it is imperative to address the equivalence of different cultural/translation variations to the fundamental psychometric specification of the original measure. Since the DLQI is the most widely used dermatology specific patient-reported outcome measure across the globe, this study investigated the unidimensionality and measurement invariance across European language variants to determine comparability of these translations to that of the original DLQI.

Polychoric (i.e. relating to the relationship between two latent variables) inter-item correlations within each country data indicated no weak relationships between items and no extremely high correlations that might indicate multicollinearity or item redundancy. Single group CFA showed unidimensionality across all the languages. Overall fit appears acceptable, but some country-specific indices suggest local misfit, including elevated RMSEA for France and elevated SRMR for Belgium, France and Turkey.

At configural level, MGCFAs may indicate that these different country translations of the DLQI conceptualize the same construct. However, qualitative evidence that respondents understand and respond to the items in a similar way

Table 4 Confirmatory factor analysis using 1-factor model by country

	All data	UK†	Belgium	Denmark	France	Germany	Hungary	Italy	Nether-lands	Norway	Poland	Russia	Spain	Turkey	
N	3408	253	218	233	107	243	204	498	203	469	233	248	249	250	
Latent variable estimates for 1-factor model															
dlqi1	0.679	0.607	0.569	0.714	0.773	0.773	0.712	0.658	0.728	0.682	0.724	0.657	0.752	0.518	
dlqi2	0.783	0.835	0.728	0.739	0.851	0.798	0.771	0.796	0.881	0.733	0.726	0.830	0.829	0.657	
dlqi3	0.854	0.840	0.729	0.874	0.918	0.803	0.880	0.830	0.913	0.860	0.866	0.846	0.863	0.872	
dlqi4	0.775	0.786	0.762	0.797	0.743	0.707	0.734	0.784	0.762	0.800	0.837	0.769	0.794	0.689	
dlqi5	0.900	0.912	0.896	0.932	0.868	0.896	0.850	0.853	0.932	0.914	0.868	0.874	0.944	0.955	
dlqi6	0.775	0.910	0.692	0.884	0.779	0.791	0.826	0.751	0.823	0.873	0.701	0.522	0.798	0.541	
dlqi7	0.727	0.783	0.756	0.787	0.862	0.746	0.847	0.770	0.699	0.731	0.674	0.699	0.822	0.596	
dlqi8	0.850	0.909	0.644	0.894	0.831	0.931	0.855	0.818	0.749	0.900	0.861	0.727	0.796	0.711	
dlqi9	0.761	0.856	0.633	0.842	0.729	0.843	0.828	0.736	0.785	0.776	0.784	0.498	0.841	0.331	
dlqi10	0.732	0.651	0.740	0.716	0.794	0.783	0.759	0.700	0.631	0.748	0.762	0.638	0.606	0.654	
Fit statistics															
RMSEA	0.063	0.005	0.045	0.062	0.108	0.039	0.075	0.061	0.048	0.064	0.092	0.088	0.000	0.072	<0.06 good fit*
Lower 90% CI RMSEA	0.058	0.000	0.008	0.039	0.076	0.000	0.051	0.047	0.011	0.050	0.071	0.068	0.000	0.051	<0.08 acceptable fit*
Upper 90% CI RMSEA	0.068	0.046	0.071	0.085	0.140	0.064	0.098	0.076	0.074	0.078	0.112	0.108	0.000	0.093	
CI (upper-lower)	0.01	0.046	0.063	0.046	0.064	0.064	0.047	0.029	0.063	0.028	0.041	0.04	0	0.042	≤ 0.06
CFI	0.994	1.000	0.991	0.996	0.985	0.998	0.994	0.995	0.998	0.996	0.991	0.987	1.000	0.983	>0.97 very good fit
Scaled CFI	0.976	0.992	0.971	0.979	0.953	0.982	0.969	0.976	0.984	0.977	0.961	0.951	0.992	0.959	>0.95 good fit*
TLI	0.993	1.000	0.989	0.994	0.981	0.998	0.992	0.994	0.997	0.995	0.988	0.983	1.000	0.978	>0.90 acceptable fit
Scaled TLI	0.970	0.990	0.962	0.973	0.940	0.976	0.960	0.969	0.979	0.971	0.951	0.936	0.990	0.947	<0.90 poor fit
NNFI	0.991	1.000	0.989	0.994	0.981	0.998	0.992	0.994	0.997	0.995	0.988	0.983	1.000	0.978	
SRMR	0.046	0.048	0.091	0.063	0.122	0.051	0.074	0.046	0.067	0.052	0.073	0.081	0.057	0.143	≤0.08*
Reliability fl															
alpha	0.899	0.900	0.812	0.911	0.885	0.905	0.904	0.901	0.895	0.910	0.903	0.870	0.864	0.803	>0.9 excellent reliability*
Omega1	0.908	0.914	0.839	0.920	0.913	0.916	0.918	0.910	0.913	0.918	0.913	0.881	0.891	0.835	>0.8 good reliability*
omega2	0.908	0.914	0.839	0.920	0.913	0.916	0.918	0.910	0.913	0.918	0.913	0.881	0.891	0.835	0.7-0.8 acceptable*
omega3	0.921	0.928	0.860	0.948	0.971	0.934	0.944	0.922	0.929	0.935	0.943	0.898	0.914	0.852	
AVE	0.618	0.664	0.518	0.675	0.667	0.655	0.653	0.596	0.633	0.649	0.614	0.514	0.654	0.454	

Values for RMSEA and SRMR < 0.08 indicate acceptable goodness-of-fit [35]

Alpha: Coefficient alpha uses the item covariance (or correlation) matrix [27]

Interpretation of Omega (ω): $\omega \approx 1$: Perfect reliability; the test or scale is highly consistent, $\omega \geq 0.9$: Very good reliability; excellent internal consistency, $0.8 \leq \omega < 0.9$: Good reliability; the test has acceptable internal consistency, $0.7 \leq \omega < 0.8$: Acceptable reliability; the scale is sufficiently consistent for general research purposes, $\omega < 0.7$: Low reliability; the scale may not be sufficiently consistent, and you may want to reconsider the instrument or conduct further analysis

Omega1: Cutoff > 0.7. Coefficient omega is a measure of composite reliability computed using the item factor loadings and uniqueness from a factor analysis i.e. total variance is the explained variance from the factor of interest plus residual error plus residual covariance [42]

Omega2: Uses total implied variance from all factors (full implied covariance matrix) i.e. reliability of a measure without controlling for another factor [42]

Omega3: McDonald's Hierarchical omega. This formula is the most conservative method in calculating coefficient omega and uses observed covariance matrix instead of model-implied covariance matrix to calculate the observed total variance. If the model fits the data well, the third coefficient omega will be similar to the other two [31, 42]

Average variance extracted is calculated from polychoric (polyserial) not Pearson correlations [18]

AVE is the average amount of variance explained by the observed variables (indicators) in a latent variable or factor [43]

†The UK version is the original language version

*Fit statistics follow COSMIN guidelines where indicated. CFA: CFI or TLI or comparable measure > 0.95 OR RMSEA < 0.06 OR SRMR < 0.08 [41]

that captures the intended construct across translations is also required, i.e. beyond what this study alone provides. At the metric (weak) level, MGCFA indicated equality of factor loadings showing that items are equally related to the same latent factor across the groups (Figure S5). However, at the scalar level (strong, equal thresholds for ordinal data), MGCFA did not suggest invariance across different

groups. Thus, comparing mean scores between the 12 different cross-cultural versions is not supported by the current results. France was not included because the responses did not fulfil the analysis criteria. This is quite evident when comparing the mean DLQI scores across countries in this dataset (Table 1), and comparing differing distributions of disease, disease severity (Figures S3 and S4), EQ VAS

Table 5 Multiple-group confirmatory factor analysis – fit measures and differences in fit measures for 12 countries

Fit measures (model)	TLI scaled	CFI scaled	RMSEA scaled	Δ CFI scaled	Δ RMSEA scaled	SRMR
Configural (baseline) ¹	0.971	0.978	0.089	NA	NA	0.067
Metric (loadings equal) ²	0.977	0.978	0.081	0.00	0.008	0.087
Scalar (thresholds equal) ³	0.962	0.948	0.104	- 0.03	- 0.023	0.079
Partial scalar (dqli7 t1) ⁴	0.980	0.977	0.076	- 0.001	- 0.005	0.068
Strict (residuals equal) ⁵	0.962	0.948	0.104	0.0	0.0	0.079
Differences in Fit Indices		df scaled	Δ RMSEA scaled	Δ CFI scaled	Δ TLI scaled	SRMR
fit.metric - fit.config ⁶		99	- 0.009	0.00	0.005	0.020
fit.scalar - fit.metric ⁷		209	0.023	- 0.03	- 0.015	-0.008
fit.partial.scalar - fit.metric		11	- 0.005	- 0.001	- 0.005	-0.019
fit.strict - fit.scalar		11	0.025	- 0.03	- 0.021	0.00

¹Good baseline fit²Metric (weak) invariance supported: Δ CFI $\leq$ 0.01 and Δ RMSEA $\leq$ 0.015³Scalar (strong) invariance not supported: Δ CFI = -0.03 ($>$ 0.01), Δ RMSEA = +0.023 ($>$ 0.015)⁴Partial Scalar invariance supported (dqli7 | threshold1 constrained: Δ CFI = -0.001 ($<$ 0.01), Δ RMSEA = -0.005 ($<$ 0.015)⁵Strict invariance not supported, as the fit measurement values were the same as for the scalar fit⁶Metric invariance (factor loadings equal) supported: Δ CFI = 0.000 ($\leq$ 0.01) and Δ RMSEA = -0.009 ($\leq$ 0.015)⁷Scalar Invariance (thresholds equal) not supported: Δ CFI = -0.030 ($>$ 0.01), Δ RMSEA = +0.023 ($>$ 0.015)

(Table 1), and plots of intercepts and thresholds (Figures S6 and S7) from country to country.

Scalar non-invariance (equal item intercepts or thresholds across groups) is needed if you want to compare latent means across languages. This means that respondents across language groups must interpret both the meaning and scaling of items similarly. A disadvantage of the MGCFA approach is that it is very strict, in the sense that it requires exact equality of parameters across groups, but in real data analysis, exact equality of measurement parameters is almost never the case [45]. This assumption of exact equivalence is not a reasonable assumption to make for measurement models. Many studies have shown that the latent factors, and associated model parameters, are not exactly equivalent across groups [11, 34].

Scalar non-invariance should not be attributed solely to differences in disease mix, severity distribution, or cultural response styles, as these factors affect observed scores, whereas measurement invariance concerns equality of item parameters across groups. Consistent with the broader psychometric literature, our findings of scalar non-invariance likely reflects some genuine item-level invariance, particularly non-invariant thresholds, rather than artefacts of sampling variation. This interpretation aligns with extensive evidence showing that full scalar invariance is rarely achieved in cross-national studies, even when configural and metric invariance are supported. Together, these findings position our results within a well-established pattern: configural and metric invariance are often attainable across countries, but full scalar invariance is uncommon, and the most parsimonious explanation is item-level DIF or non-equivalent thresholds, rather than sample composition effects. For example a study examining the SWLS across 65

nations found that while configural and metric invariance were upheld, full scalar invariance was only achieved across gender identities and age groups [46]. Research using the Quality of Life Enjoyment and Satisfaction Questionnaire-Short Form (Q-LES-Q-SF) across ten countries revealed that 12 out of 14 items were non-invariant, indicating that full scalar invariance was not established. This suggested that individuals from different countries may interpret items differently, affecting the comparability of scores [47]. A review of 27 articles measuring invariance of the Satisfaction with Life Scale (SWLS) found that studies generally supported a unidimensional structure (configural), but less commonly supported were equivalent factor loadings (metric), whereas scalar or strict measurement invariance was supported in only 1 of the 11 cultural analyses [48]. A systematic review of 75 peer-reviewed articles using personality measures valid for different populations found that none of these studies achieved scalar or strict measurement invariance across cultural/ethnic groups [49]. Together these findings suggest that while configural and metric invariance are often achievable in cross-nationality studies, full scalar invariance is less common. We are also carrying out item-level DIF analyses and investigating partial metric invariance based on modification indices but these have not improved the models.

Response heterogeneity refers to systematic differences in how individuals or subgroups interpret, understand, or use questionnaire response options, even when they have the same underlying health status. This manifests as DIF, differential interpretations of items across age, sex, culture, severity, or clinical characteristics or latent subgroups showing different response patterns [50, 51]. Response heterogeneity can occur within countries/cultures (e.g. depending on gender, age, level of education, level of literacy) [52]. DIF

analysis using IRT, ordinal logistic regression DIF, multi-level DIF models, mixture IRT, unsupervised or tree based IRT or partial measurement invariance CFA / MG-CFA can all be used to investigate response heterogeneity. Many methods have been proposed to manage or correct response heterogeneity including using DIF-adjusted scoring (e.g. IRT-based recalibration of item parameters), accepting partial invariance modelling in CFA / MG-CFA (allowing some parameters to vary), use of latent subgroup-specific scoring (e.g. MixIRT), improving questionnaire design and cultural adaptation, and developing core outcome sets or harmonised instruments [53].

Additionally, the sensitivity of goodness of fit indexes and of the chi-square difference statistics to lack of intercept invariance has been demonstrated to be affected by the interaction between the proportion of invariance and pattern of invariance, main effect of pattern of invariance, and ratio of sample sizes [54].

Known-group analysis is a method used in psychometrics and validation studies to assess construct validity—specifically, whether an instrument has the ability to differentiate differences between distinct (independent) groups within the construct being measured. Groups are generally defined using another independent measure [55], for example the DLQI score meaning bands of Hongbo et al. [38]. Group differences are then determined using a statistical test and the effect size can also be determined [56]. Known group validity analysis of DLQI total score by severity was significant between groups for all country data, with a wide range of effects sizes from none to large for changes in severity from mild to moderate severity, and moderate to severe, and broad confidence intervals. This may reflect lack of uniformity in physician assessed severity across countries, and we were unable to evaluate this further.

Criticism of the DLQI score meaning banding centres around concern that questions considered by respondents as “not relevant” may result in a score sum not reaching a level that a guideline document may deem necessary to consider a specific therapy. However, clinical judgement is always required in the interpretation of clinical scores in specific patient situations [57].

The strengths of this work are that essential unidimensionality and configural (baseline) and metric invariance (loadings equal) and partial scalar invariance have been demonstrated for the DLQI across these European countries. However full scalar (thresholds equal) and means invariance was not demonstrated likely reflecting some genuine item-level non-invariance, particularly non-invariant thresholds, rather than artefacts of sampling variation. This could be further investigated with Bayesian approximate measurement invariance methods [47, 58], alignment methods and item-level DIF analyses. Some confidence should also be

gained from the demonstration of known group analysis across the countries, although the physician base assessment of severity may not be uniform. Additionally, previously reported [59, 60] differential item functioning (DIF) and evidence of good longitudinal response of the DLQI (ability to detect change over time) needs to be investigated within individual countries of this dataset.

Limitations of this study include a lack of historic records relating to cognitive debriefing of translations and that could be a plausible source of non-invariance. Additionally, the performance of European language versions of the DLQI used in this study may not represent the performance in non-European cultural contexts, particularly Asian, Middle Eastern or African, which have very different cultural attitudes to those captured in the European context.

These results support evidence for a dominant general factor (essential unidimensionality) and metric invariance of the DLQI, but not valid cross-country comparison of means across these different European languages. The construct validity of the DLQI and known group validity was demonstrated with the anchor of physician assessed severity across all the translations. This analysis provides further information relevant to researchers involved in international studies using the DLQI.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00403-026-04877-0>.

Acknowledgements We wish to thank the physicians and patients across Europe for their contribution to the study data generation.

Author contributions CRediT: Conceptualization AF, SS, JJ, FA; Data Curation JK, JJ, FA; Formal Analysis JJ; Investigation JJ, JK; Methodology JJ, AF, SS, JK, FD, FA; Project Administration/Resources/Supervision AF, SS, FD, JK; Validation SS, FA, JK; Writing – Original Draft JJ; Writing – Review & Editing JJ, AF, SS, JK, FD, FA.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests Andrew Y Finlay is joint copyright owner of the DLQI. Cardiff University receives royalties from some use of the DLQI: AYF receives a proportion of these under standard university policy. Sam Salek has received an unrestricted educational grant from GSK, is a consultant for Novo Nordisk and produces educational materials for Abbvie. Faraz Ali has received honorariums from Abbvie, Janssen, LEO pharmaceuticals, Lilly pharmaceuticals, L’Oreal, Novartis and UCB. Faraz Ali and Jeffrey Johns’ department receives income from royalties from the DLQI and related instruments.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate

if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Finlay AY, Khan GK (1994) Dermatology Life Quality Index (DLQI)—a simple practical measure for routine clinical use. *Clin Exp Dermatol* 19(3):210–216. <https://doi.org/10.1111/j.1365-2230.1994.tb01167.x>
- Cardiff University (2024) Dermatology Life Quality Index. <https://www.cardiff.ac.uk/medicine/resources/quality-of-life-questionnaires/dermatology-life-quality-index>. Accessed 18 Jun 2026
- Singh R, Finlay AY Dermatology Life Quality Index use in skin disease guidelines and registries worldwide. *J Eur Acad Dermatol Venereol*. 34(12):e822–e4. <https://doi.org/10.1111/jdv.16701>
- Vyas J, Johns JR, Ali FM, Singh RK, Ingram JR, Salek S et al (2023) A systematic review of 457 randomised controlled trials using the Dermatology Life Quality Index: experience in 68 diseases and 42 countries. *Br J Dermatol* 190(3):315–339. <https://doi.org/10.1093/bjd/ljad079>
- Johns JR, Vyas J, Ali FM, Ingram JR, Salek S, Finlay AY (2024) The Dermatology Life Quality Index as the primary outcome in randomized clinical trials: a systematic review. *Br J Dermatol* 191(4):497–507. <https://doi.org/10.1093/bjd/ljae228>
- Vyas J, Johns JR, Abdelrazik Y, Ali FM, Ingram JR, Salek S et al (2024) The Dermatology Life Quality Index (DLQI) used as the benchmark in validation of 101 quality-of-life instruments: A systematic review. *J Eur Acad Dermatol Venereol* 39(3):631–679. <https://doi.org/10.1111/jdv.20321>
- Johns J, Salek S, Ali FM, Dalgard F, Kupfer J, Finlay AY (2026) Confirmation of unidimensionality of the Dermatology Life Quality Index (DLQI) using a multinational 3,408 patient dataset. *J Patient Rep Outcomes Press*. <https://doi.org/10.1186/s41687-026-01025-0>
- Vyas J, Johns JR, Ali FM, Ingram JR, Salek S, Finlay AY (2024) A Systematic Review of 207 Studies Describing Validation Aspects of the Dermatology Life Quality Index. *Acta Derm Venereol* 104:adv41120. <https://doi.org/10.2340/actadv.v104.41120>
- Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL et al (2010) The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 63(7):737–745. <https://doi.org/10.1016/j.jclinepi.2010.02.006>
- Liu X, Cao P, Lai X, Wen J, Yang Y (2023) Assessing Essential Unidimensionality of Scales and Structural Coefficient Bias. *Educ Psychol Meas* 83(1):28–47. <https://doi.org/10.1177/00131644221075580>
- Millsap RE (2011) Statistical approaches to measurement invariance. Routledge, New York
- Yamazaki Y, Kayes DC (2004) An Experiential Approach to Cross-Cultural Learning: A Review and Integration of Competencies for Successful Expatriate Adaptation. *Acad Manag Learn Educ* 3(4):362–379
- Beaton DE, Bombardier C, Guillemin F, Ferraz MB (2000) Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine* 25(24):3186–3191
- Vyas J, Johns JR, Trivedi A, Ali FM, Ingram JR, Salek S et al (2025) Systematic review of the use of the Dermatology Life Quality Index in routine clinical practice: evidence from 287 articles across 56 countries. *Clin Exp Dermatol* 50(12):2456–2465. <https://doi.org/10.1093/ced/llaf343>
- Dalgard FJ, Gieler U, Tomas-Aragones L, Lien L, Poot F, Jemec GBE et al (2015) The psychological burden of skin diseases: a cross-sectional multicenter study among dermatological outpatients in 13 European countries. *J Invest Dermatol* 135(4):984–991. <https://doi.org/10.1038/jid.2014.530>
- Kamudoni P, Johns N, Salek S (2018) Living with Chronic Disease: Measuring Important Patient-Reported Outcomes. Adis, Springer Nature, Singapore
- Kiwanuka F, Kopra J, Sak-Dankosky N, Nanyonga RC, Kvist T (2022) Polychoric Correlation With Ordinal Data in Nursing Research. *Nurs Res* 71(6):469–476. <https://doi.org/10.1097/nnr.0000000000000614>
- DeVellis RF (2017) Scale Development: Theory and Applications, 4th edn. Sage, Thousand Oaks, CA
- Brown TA (2015) Confirmatory Factor Analysis for Applied Research Second Edition. Guilford, New York
- Kline RB (2016) Principles and Practice of Structural Equation Modeling, 4th edn. Guilford Press, New York
- Schreiber JB, Nora A, Stage FK, Barlow EA, King J (2006) Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review. *J Educ Res* 99(6):323–338. <https://doi.org/10.3200/JOER.99.6.323-338>
- Muthén BO, du Toit SHC, Spisic D (1997) Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Retrieved from https://www.statmodel.com/download/Article_075.pdf
- Millsap RE, Yun-Tein J (2004) Assessing factorial invariance in ordered-categorical measures. *Multivar Behav Res* 39(3):479–515
- Holgado-Tello FP, Chacón-Moscó S, Barbero-García I, Vila-Abad E (2010) Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Qual Quant* 44(1):153–166. <https://doi.org/10.1007/s11135-008-9190-y>
- Muthén BO, Bollen KA, Long JS (1993) Goodness of fit with categorical and other non-normal variables. *Testing Structural Equation Models*. A: Sage, Newbury Park, CA: Sage; pp 205–243
- Cronbach LJ, Warrington WG (1951) Time-limit tests: estimating their reliability and degree of speeding. *Psychometrika* 16(2):167–188. <https://doi.org/10.1007/BF02289113>
- Cronbach LJ, Meehl PE (1955) Construct validity in psychological tests. *Psychol Bull* 52(4):281–302. <https://doi.org/10.1037/h0040957>
- Raykov T (2001) Estimation of congeneric scale reliability using covariance structure analysis with nonlinear constraints. *Br J Math Stat Psychol* 54(Pt 2):315–323. <https://doi.org/10.1348/000711001159582>
- Bentler PM (1972) A lower-bound method for the dimension-free measurement of internal consistency. *Soc Sci Res* 1(4):343–357. [https://doi.org/10.1016/0049-089X\(72\)90082-8](https://doi.org/10.1016/0049-089X(72)90082-8)
- Bentler PM, Alpha D-F (2009) Model-Based Internal Consistency Reliability. *Psychometrika* 74(1):137–143. <https://doi.org/10.1007/s11336-008-9100-1>
- McDonald RP (1970) The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *Br J Math Stat Psychol* 23(1):1–21. <https://doi.org/10.1111/j.2044-8317.1970.tb00432.x>
- Bowen NK, Masa RD (2015) Conducting measurement invariance tests with ordinal data: a guide for social work researchers. *Univ of Chicago*, pp 229–249

33. Steenkamp J-BEM, Baumgartner H (1998) Assessing Measurement Invariance in Cross-National Consumer Research. *J Consum Res* 25(1):78–90. <https://doi.org/10.1086/209528>
34. Vandenberg RJ, Lance CE (2000) A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organ Res Methods* 3(1):4–70. <https://doi.org/10.1177/109442810031002>
35. Awang Z (2012) A handbook on SEM. Structural equation modeling: SEM Using AMOS Graphic, 5th edn. UiTM, Kota Baru
36. Jonckheere AR (1954) A distribution-free k-sample test against ordered alternatives. *Biometrika* 41(1/2):133–145
37. Terpstra TJ (1952) The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. *Indagationes Math* 14(3):327–333
38. Hongbo Y, Thomas CL, Harrison MA, Salek MS, Finlay AY (2005) Translating the science of quality of life into practice: What do dermatology life quality index scores mean? *J Invest Dermatol* 125(4):659–664
39. Shi D, Maydeu-Olivares A, Rosseel Y (2020) Assessing Fit in Ordinal Factor Analysis Models: SRMR vs. RMSEA. *Struct Equ Model* 27(1):1–15. <https://doi.org/10.1080/10705511.2019.1611434>
40. Maydeu-Olivares A, Joe H (2014) Assessing Approximate Fit in Categorical Data Analysis. *Multivar Behav Res* 49(4):305–328. <https://doi.org/10.1080/00273171.2014.911075>
41. Prinsen CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HCW et al (2018) COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res* 27(5):1147–1157. <https://doi.org/10.1007/s11136-018-1798-3>
42. Green SB, Yang Y (2009) Reliability of Summed Item Scores Using Structural Equation Modeling: An Alternative to Coefficient Alpha. *Psychometrika* 74(1):155–167. <https://doi.org/10.1007/S11336-008-9099-3>
43. Fornell C, Larcker DF (1981) Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. *J Mark Res* 18(1):39–50. <https://doi.org/10.2307/3151312>
44. Chen FF (2007) Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Struct Equ Model* 14:464–504. <https://doi.org/10.1080/10705510701301834>
45. Davidov E, Meuleman B, Cieciuch J, Schmidt P, Billiet J (2014) Measurement Equivalence in Cross-National Research. *Annu Rev Sociol* 40:55–75. <https://doi.org/10.1146/annurev-soc-071913-043137>
46. Swami V, Stieger S, Voracek M, Aavik T, Abdollahpour Ranjbar H, Adebayo SO et al (2025) Life satisfaction around the world: Measurement invariance of the Satisfaction With Life Scale (SWLS) across 65 nations, 40 languages, gender identities, and age groups. *PLoS ONE* 20(1):e0313107. <https://doi.org/10.1371/journal.pone.0313107>
47. Bagheri Z, Chamanpara P, Jafari P, Balhara YPS, Arya S, Ransing R et al (2022) Cross-cultural measurement invariance of the Quality of Life Enjoyment and Satisfaction Questionnaire-Short form across ten countries: the application of Bayesian approximate measurement invariance. *BMC Psychol* 24(10):160
48. Emerson SD, Guhn M, Gadermann AM (2017) Measurement invariance of the Satisfaction with Life Scale: reviewing three decades of research. *Qual Life Res* 26(8):2251–2264. <https://doi.org/10.1007/s11136-017-1552-2>
49. Dong Y, Dumas D (2020) Are personality measures valid for different populations? A systematic review of measurement invariance across cultures, gender, and age. *Pers Individ Dif* 160:109956. <https://doi.org/10.1016/j.paid.2020.109956>
50. Penton H, Dayson C, Hulme C, Young T (2022) An Investigation of Age-Related Differential Item Functioning in the EQ-5D-5L Using Item Response Theory and Logistic Regression. *Value Health* 25(9):1566–1574
51. Kaptur DC, Liu Y, Kaptur B, Peterman N, Zhang J, Kern JL et al (2025) Examining differential item functioning in self-reported health survey data: via multilevel modeling. *Qual Life Res* 34(6):1569–1577. <https://doi.org/10.1007/s11136-025-03936-9>
52. Valentine N, Verdes-Tennant E, Bonsel G (2015) Health systems' responsiveness and reporting behaviour: Multilevel analysis of the influence of individual-level factors in 64 countries. *Soc Sci Med* 138:152–160. <https://doi.org/10.1016/j.socscimed.2015.04.022>
53. Teresi JA (2006) Overview of Quantitative Measurement Methods: Equivalence, Invariance, and Differential Item Functioning in Health Applications. *Med Care* 44(11):S39–S49. <https://doi.org/10.1097/01.mlr.0000245452.48613.45>
54. Chen FF (2007) Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Struct Equ Model* 14(3):464–504. <https://doi.org/10.1080/10705510701301834>
55. Tabachnick BG, Fidell LS (2019) Using multivariate statistics, 7th edn. Pearson, London
56. Sullivan GM, Feinn R (2012) Using Effect Size-or Why the P Value Is Not Enough. *J Grad Med Educ* 4(3):279–282. <https://doi.org/10.4300/JGME-D-12-00156.1>
57. Finlay AY, Sampogna F (2018) What do scores mean? Informed interpretation and clinical judgement are needed. *Br J Dermatol* 179(5):1021–1022. <https://doi.org/10.1111/bjd.17028>
58. Van de Vijver FRJ, Avisati F, Eid M, Fox J-P, Le Donné N, Lek K et al (2018) Invariance analyses in large-scale studies: OECD Education Working Papers No. 201. OECD; p. 110
59. Nijsten T, Meads DM, de Korte J, Sampogna F, Gelfand JM, Ongena K et al (2007) Cross-cultural inequivalence of dermatology-specific health-related quality of life instruments in psoriasis patients. *J Invest Dermatol* 127(10):2315–2322. <https://doi.org/10.1038/jid.2008.185>
60. Jorge MFS, Sousa TD, Pollo CF, Paiva BSR, Ianhez M, Boza JC et al (2020) Dimensionality and psychometric analysis of DLQI in a Brazilian population. *Health Qual Life Outcomes* 18(1):268. <https://doi.org/10.1186/s12955-020-01523-9>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.