Research Article

Inter-Rater Reliability And Agreement Of The Diagnostic Assessment Scale For Kushtha In Papulosquamous Skin Disease: A Cross-Sectional Two-Rater Study

Arun Kumar M[1]*, Nagaraj S[2], Saritha T[3]

[1]PhD Scholar (Preliminary), Department of Roga Nidana evum Vikriti Vijnana, SDMCAH&RC, UDUPI,574118, Karnataka, India. Email: arunsdm@sdmayurvedacollegeudupi.in. ORCID: 0000-0003-0874-463X.
[2]Professor and Head, Department of Roga Nidana evum Vikriti Vijnana, SDMCAH&RC, UDUPI,574118, Karnataka, India. Karnataka, India. Email: consultantdr@rediffmail.com. ORCID-ID:0000-0002-5098-332.
[3]Assistant Professor, Department of Roga Nidana evum Vikriti Vijnana, SDMCAH&RC, UDUPI,574118, Karnataka, India. Email: sarithathonse83@gmail.com ORCID:0000-0002-5290-8075

*Corresponding author: Dr Arun Kumar M, PhD Scholar (Preliminary), Department of Roga Nidana evum Vikriti Vijnana, SDMCAH & RC, UDUPI,574118, Karnataka, India.
Email: arunsdm@sdmayurvedacollegeudupi.in. 0000-0003-0874-463X

Abstract

Background and objectives: The Diagnostic Assessment Scale for Kushtha (DASK) is a newly developed 26-item ordinal instrument that renders the classical threefold Ayurvedic examination as a severity score and a dosha attribution. No reliability estimate has been published. This study estimated its inter-rater reliability, agreement, and the measurement error attaching to an individual score.

Methods: Sixty consecutive patients with consultant-confirmed papulosquamous disease (43 psoriasis, 17 lichen planus) were each assessed independently by two postgraduate-qualified Ayurvedic physicians in a single clinical session, giving a fully crossed design of 3,120 item scores. Reliability was estimated as the intraclass correlation coefficient ICC(2,1); agreement as the standard error of measurement (SEM), minimal detectable change (MDC_{95}) and Bland–Altman limits; and item-level agreement as quadratic weighted kappa with Gwet's AC_2. Internal consistency and categorical agreement were also assessed, following the GRRAS guidelines.

Results: No data were missing. ICC(2,1) was 0.952 (95% confidence interval 0.92–0.97) for the total score and 0.945, 0.920 and 0.920 for the Vata, Pitta and Kapha subscales. Between-patient variance accounted for 95.2% of total variance and the rater effect for 0.0–0.6%. SEM was 2.60 points and MDC_{95} 7.21 points on the 0–104 scale, and all 26 items reached substantial or almost-perfect agreement (weighted kappa 0.732–0.963). The categorical outputs were less reliable than the scores generating them: severity band kappa 0.800 and dosha attribution kappa 0.640 (0.37–0.85), every disagreement occurring at a cut-point or a narrow margin. Kapha was reproduced consistently yet was not internally coherent (alpha 0.43 and 0.42).

Conclusion: DASK measures the severity of papulosquamous Kushtha reliably enough for group comparison and, against a threshold of 8 points, for individual monitoring. Agreement on its categorical outputs was lower than on the continuous scores from which they are derived, and the Kapha subscale was reproduced consistently between raters without cohering internally.

Keywords: Ayurveda, Kushtha, Psoriasis, Lichen Planus, Inter-Rater Reliability, Intraclass Correlation, Minimal Detectable Change, Weighted Kappa.

1. Introduction

Classical Ayurvedic assessment of Kushtha rests on narrative clinical description. The tradition carries considerable diagnostic subtlety but does not yield a number, and without a number a patient cannot be compared with himself three months later, one clinic's caseload cannot be compared with another's, and Ayurvedic dermatology cannot enter the evidential currency of controlled trials. The Diagnostic Assessment Scale for Kushtha (DASK) was developed to address this gap, mapping the three classical examination modalities onto 26 ordinally anchored items and returning both a severity total and a dosha attribution.

An instrument of this kind is of value only if two competently trained clinicians examining the same patient arrive at the same number. Inter-rater reliability and inter-rater agreement bound everything the instrument can subsequently be used for; an instrument on which two raters differ by eight points cannot detect a six-point treatment effect, however large the trial. The two properties are distinct and are often conflated. Reliability describes how well an instrument distinguishes patients from one another despite measurement error and, being a ratio of true-score variance to total variance, depends on how heterogeneous the sample is. Agreement describes how close two raters' scores are in the units of the scale itself and does not depend on heterogeneity. A study reporting only the first has not told the clinician what he needs to know.[1] Both are reported here.

No reliability estimate for DASK has been published, so there is no prior figure against which the present estimates can be benchmarked. The wider dermatological literature supplies a qualitative calibration. Reported inter-rater reliability of established severity indices such as the Psoriasis Area and Severity Index varies appreciably with rater training and case mix, and the items that historically degrade such indices are those calling for a graded judgement of a continuous physical property rather than of presence or absence. DASK contains several of these. A further hazard applies to instruments built partly on interview: if two

raters hear the same answer from the same patient, their scores are two transcriptions of a single utterance rather than two independent observations. Eight of the 26 DASK items are interview items, so this concern is structural and was examined empirically.

The primary objective was to estimate the inter-rater reliability of the DASK total score and of each dosha subscale as ICC(2,1) with 95% confidence intervals (CIs). Secondary objectives were to quantify agreement in the units of the scale, to locate disagreement at item level, to estimate the reliability of the two categorical outputs that reach the clinical record, and to assess internal consistency as a check on whether each subscale measures a single construct.

2. Materials and Methods

2.1. Study design and setting

This was a cross-sectional, fully crossed, two-rater reliability and agreement study conducted in the dermatology outpatient and inpatient services of SDM College of Ayurveda Hospital and Research Centre (SDMCAH&RC), Udupi, Karnataka, India. Every patient was assessed by both raters, giving a complete 60 × 2 subjects-by-raters matrix with no missing cells. A fully crossed design permits the rater main effect to be separated from residual error, which is what makes an absolute-agreement coefficient and a variance decomposition available. Reporting follows the Guidelines for Reporting Reliability and Agreement Studies (GRRAS).[2]

2.2. Participants

Consecutive patients attending the service were screened against the eligibility criteria and invited. 60 were enrolled and completed both assessments. Consecutive sampling was preferred to convenience selection because reliability coefficients are inflated by sample heterogeneity, and a sample assembled by deliberately selecting clear-cut cases at the extremes of severity would return a flattering coefficient that would not survive contact with a routine caseload.

Patients were eligible if they had a consultant-confirmed papulosquamous dermatological diagnosis (psoriasis, lichen planus or a related condition), were aged 18 years or above, and were willing and able to give informed consent and to undergo two independent examinations within a single clinical session. Patients were excluded if they were on active systemic therapy likely to alter lesion morphology between the two assessments, if lesions were confined to sites the patient declined to expose to both raters, or if the condition changed visibly between the two assessments. The reference dermatological diagnosis was established by the attending consultant, not by either rater.

2.3. The instrument

DASK comprises 26 items, each graded 0–4 against written anchor descriptors and distributed across three examination sections: Darshana Pariksha (inspection, items 1–12), Sparshana Pariksha (palpation, items 13–18) and Prashna Pariksha (interview, items 19–26). The items also map onto three dosha subscales, and this mapping does not follow the section structure (Table 1).

Three features of the scoring model bear on the findings. First, the dosha attribution is made on percentages of each subscale maximum rather than on raw sums (Vata % = Vata ÷ 44, Pitta % = Pitta ÷ 28, Kapha % = Kapha ÷ 32), the highest percentage naming the dominant dosha; the normalisation exists because the subscales contain unequal numbers of items. Second, item 25 (Sthairya) is reverse anchored, grade 4 denoting a lesion static for six months or more and grade 0 a rapidly spreading lesion. Third, severity is banded from the total score as Mild 0–34, Moderate 35–69 and Severe 70 or above.

2.4. Raters and rating procedure

Two Ayurvedic physicians, both holding postgraduate qualifications in Roga Nidana, served as raters. Rater 1, the principal investigator, developed the DASK anchor definitions and trained Rater 2 in their application at a calibration session held before data collection began. That the two raters stand in a trainer–trainee relationship is stated plainly because it bears on generalisability: a rater trained directly by the developer of an instrument is likely to agree with that developer more closely than an independent physician working from the manual alone, so the coefficients reported here are best read as an upper bound. The inference target is nevertheless the population of trained Ayurvedic physicians who might reasonably administer DASK after the training the manual specifies, which is why the rater factor is treated as random.

The protocol specified that both raters assess the same patient during the same clinical session, with an interval between the two examinations of approximately 120 minutes; that Rater 1 complete and seal the DASK form before Rater 2 began; that Rater 2 not see Rater 1's form at any point; that the raters not communicate before or during assessment; that the patient be in the same position and location for both examinations; and that lighting and ambient temperature be identical. Item 15 (Shaitya) requires an infrared or contact thermometer.

2.5. Sample size

Sample size for a reliability study is governed by the precision of the resulting interval rather than by power against a null hypothesis. Bonett's formula[3] with an anticipated reliability of 0.90, two raters and a target CI half-width of 0.05 returns a requirement of 57 patients. A lower anticipated reliability would require a larger sample, so this assumption is not conservative and the achieved precision depends on the true value being close to it. Sixty were recruited, giving 60 × 26 × 2 = 3,120 item scores.

2.6. Statistical analysis

The primary statistic was ICC(2,1) in the notation of Shrout and Fleiss,[4] that is two-way random effects, absolute agreement, single measurement, equivalently ICC(A,1) of McGraw and Wong.[5] Absolute agreement rather than consistency was required because DASK scores are read against fixed severity bands: a rater scoring every patient three points higher than his colleague preserves rank order perfectly and would still misclassify patients at the band boundaries. ICC(3,1) was reported as a diagnostic, the gap between the two indexing systematic rater bias, and ICC(2,k) as the coefficient relevant if two raters' scores were averaged. CIs used the exact F-distribution method.[5] Interpretation followed Koo and Li:[6] 0.90 or above excellent, 0.75–0.89 good, 0.50–0.74 moderate, below 0.50 poor.

Because the ICC contains sample variance in its denominator it cannot state how many points of error attach to an individual patient's score. The standard error of measurement, $SEM = SD_{pooled} \times \sqrt{(1 - ICC)}$, and the minimal detectable change, $MDC_{95} = 1.96 \times \sqrt{2} \times SEM$, were computed for this purpose.[7,8] For each scale the difference between raters was plotted against their mean (Bland–Altman),[9,10] with the mean difference and its 95% CI, the 95% limits of agreement, and a test for proportional bias by regression of the difference on the mean. Normality of the difference distributions was checked using the Shapiro–Wilk test[11] and the Wilcoxon signed-rank test was reported alongside the paired t test.

Item-level agreement was quantified using Cohen's quadratic weighted kappa (κ_w)[12] over the full 0–4 category space, quadratic weighting being the standard choice for ordinal data because it credits near misses and penalises large disagreements disproportionately. Weighted kappa was preferred to correlation, which measures association rather than agreement. CIs were obtained by bootstrap resampling of patients (3,000 replicates, fixed seed) and interpretation followed Landis and Koch.[13] Because kappa is depressed when marginal distributions are skewed,[14] Gwet's AC_2 with quadratic weights[15,16] was computed for every item as a paradox-resistant comparator.

Cronbach's alpha[17] was computed separately for each rater, with McDonald's omega alongside because alpha assumes essentially tau-equivalent items, an assumption a clinical observation scale rarely satisfies.[18] Corrected item-total correlations were used to locate items failing to cohere with the construct. Agreement on the severity band and on the dosha attribution was quantified using Cohen's kappa with bootstrap CIs, and marginal homogeneity was tested using the Stuart–Maxwell test.[19] Both categorical outputs were derived by applying the manual's algorithm to each rater's own 26 item scores; agreement on the diagnosis labels as recorded is reported alongside.

Data were entered and managed in Microsoft Excel [version], which was also used to compute Gwet's AC_2 with quadratic weights, the Bland–Altman difference and mean variables, and the standard error of measurement and minimal detectable change. Intraclass correlation coefficients with their confidence intervals, Cronbach's alpha, corrected item-total correlations, the analysis of variance mean squares, Cohen's kappa for the categorical outputs, the Shapiro–Wilk test and the paired t and Wilcoxon signed-rank tests were computed in IBM SPSS Statistics (IBM Corp., Armonk, NY, USA). Python 3.12 was used for the bootstrap confidence intervals, which resample at the patient level because patients rather than individual observations are the independent units, for the Stuart–Maxwell test of marginal homogeneity, and for McDonald's omega. Bootstrap procedures used a fixed seed, so every interval is exactly reproducible, and each principal estimator was verified against an independent implementation written from its defining formula, agreeing to at least six decimal places.

Before any reliability statistic was computed the dataset was audited for completeness, permitted range, key alignment across the two workbooks, internal arithmetic and demographic consistency.

3. Results

3.1. Sample and data completeness

Sixty patients were assessed by both raters (mean age 46.2 years, standard deviation 13.7, range 18–70; 33 male and 27 female; 43 psoriasis and 17 lichen planus). All 3,120 item scores were present, all were integers within the permitted 0–4 range, and each rater's recorded subscale and total scores reconciled exactly with the sums of the constituent items in all 60 records. Patient identifier, age, sex and dermatological diagnosis were identical across the two files.

Two features of the score distribution constrain what follows (Table 2). The total score spanned 18 to 76 across both raters, about 56% of the instrument's 0–104 design range, and the two raters placed only one and two patients respectively in the Severe band, so the severity classification is very nearly a two-category problem in this sample and the Severe band is barely tested. Separately,

17 of the 26 items exceeded a 15% floor and 8 exceeded a 15% ceiling (Supplementary Table S6), which is the operating condition under which the kappa paradox becomes active. Item 15 (Shaitya) was the most extreme, sitting at floor for the large majority of patients.

3.2. Independence of the interview items

Had the two raters been sharing a single set of interview responses, the eight Prashna items would have shown near-identical agreement across all patients and would have stood clearly apart from the observational items. They did not. Exact agreement on the Prashna items ranged from 65.0% to 95.0% with κ_w from 0.81 to 0.96, spanning rather than sitting above the range occupied by the Darshana and Sparshana items, and the least reproducible item in the whole instrument (Gaurava, 65.0% exact) was itself an interview item. No item in the instrument reached perfect agreement.

3.3. Reliability, variance components and measurement error

The total DASK score achieved ICC(2,1) = 0.952 (95% CI 0.92 to 0.97), the entire interval lying above 0.90 (Table 3). The three subscales followed closely: Vata 0.945 (0.91 to 0.97), Pitta 0.920 (0.87 to 0.95) and Kapha 0.920 (0.87 to 0.95). All three lower confidence limits exceeded 0.85, so the subscales may be reported individually.

For the total score, 95.2% of observed variance was genuine between-patient variance and 4.8% residual error. On Vata, Pitta and the total score the estimated rater variance component was negative and was truncated to zero, meaning the observed difference between raters was smaller than residual noise alone would predict; only Kapha carried a positive rater component, at 0.6%. Disagreement between these observers was therefore predominantly occasion-specific rather than a fixed difference in how the anchors were read, the single exception being the small Kapha offset described below.

SEM on the total score was 2.60 points and MDC_{95} 7.21 points. A patient whose total score falls by five points has therefore not measurably improved, whereas a fall of eight points or more can be distinguished from measurement noise with 95% confidence. MDC_{95} is 6.9% of the nominal range and 12.4% of the 58-point range actually observed, the latter being the more relevant denominator. It consumed 13.9% of Vata's observed range, 17.0% of Pitta's and 14.2% of Kapha's, so the total score remains the appropriate outcome for individual monitoring.

3.4. Agreement

On three of the four scales the raters were statistically indistinguishable (Table 4, Figure 1). The total score showed a mean difference of +0.28 points (95% CI −0.67 to 1.24, p = 0.555), Vata +0.15 (p = 0.562) and Pitta −0.18 (p = 0.462). Kapha was the exception: Rater 1 scored it 0.32 points higher on average (95% CI 0.05 to 0.59, p = 0.023; Wilcoxon p = 0.027). The effect is well inside that subscale's MDC_{95} of 2.13 points and is of no consequence for an individual patient, but it is the source of the ICC(2,1) to ICC(3,1) gap and it acts on one side of the comparison that decides the dosha attribution.

The proportional bias test was non-significant on every scale (all $p \geq 0.18$) and limits of agreement for the total score ran from −6.96 to +7.53 points. Difference distributions departed from normality on every scale, expected because differences between two closely agreeing raters on an integer scale pile up at zero; the parametric limits should accordingly be read as indicative, though parametric and non-parametric tests agreed throughout as to which biases were significant.

3.5. Item-level agreement

Every item reached substantial or almost-perfect agreement on the Landis and Koch scale, with κ_w ranging from 0.732 to 0.963, 20 of the 26 items in the almost-perfect band and none below 0.61 (Table 5, Figure 2). The largest mean difference between raters on any item was 0.33 of a grade point. The strongest items were Kleda-2 (0.963), Shula (0.962) and Kandu (0.927); the weakest were Shvaitya (0.732), Shosha (0.770) and Shyava (0.775), all still substantial. Shosha and Shyava each require a graded judgement of a continuous visual property; Shvaitya is a floor-effect artefact and is considered separately below.

The floor effect must be kept in view when reading the lowest kappas. Shvaitya carried the lowest κ_w in the instrument, yet the raters agreed exactly on 88.3% of patients and its AC_2 was 0.98; the same pattern held for Shaitya (κ_w 0.783, exact agreement 96.7%, AC_2 1.00). These are items on which almost every patient scored zero, so that chance agreement is high and kappa correspondingly deflated, and where the two coefficients diverge this sharply the item is uninformative rather than unreliable. Kappa and AC_2 agreed closely for items with well-spread distributions and diverged only for near-floor items, and no item was weak on both coefficients.

3.6. Internal consistency

The full 26-item scale was internally consistent for both raters (alpha 0.863 and 0.863; omega 0.89 and 0.89) and Pitta was consistent for both (0.836 and 0.810). Vata was marginal at 0.710 and 0.711, sitting at the conventional 0.70 boundary.[20] The

Kapha subscale was not internally coherent (alpha 0.426 and 0.418; omega 0.49 and 0.51), meeting neither the 0.70 usually cited for group comparison nor the 0.60 sometimes accepted for exploratory work (Table 6).

Kapha's inter-rater reliability was excellent (ICC 0.920) while its internal consistency was poor. These findings answer different questions and are not contradictory: both raters were measuring the same things in the same way, but those things do not appear to be eight manifestations of one underlying construct.

Corrected item-total correlations showed the incoherence was not confined to Kapha. Seven items correlated below 0.30 with the remainder of the scale for at least one rater and six did so for both; of those six, Aruna (item 6), Harsha (item 19) and Kandu (item 24) had near-zero or negative corrected correlations for both raters. Shvaitya, Shaitya and Sneha also correlated weakly throughout, consistent with their near-floor distributions.

3.7. Agreement on the categorical outputs

Severity band classification agreed substantially (kappa 0.800, 95% CI 0.63 to 0.93; 90.0% exact agreement), with 6 disagreements in 60, every one of them at a band boundary (Table 6). The Stuart–Maxwell test found no evidence of marginal asymmetry (χ^2 = 2.80, df = 2, p = 0.25), so the disagreements were symmetric rather than one rater grading consistently more severely.

Dosha attribution was the weakest output (kappa 0.640, 95% CI 0.37 to 0.85), the raters agreeing on 52 of 60 patients. The CI reaches down into the fair band, and 8 patients would receive a different dosha attribution, and potentially a different treatment, depending on which of two equally qualified physicians examined them. Agreement on the labels as actually recorded was similar (kappa 0.752), so the finding is not an artefact of recomputation.

Among the 52 patients on whom the raters agreed, the margin between the highest and second-highest dosha percentage averaged well above ten percentage points, whereas among the 8 discordant patients it was typically under five points and in several cases a single percentage point. No patient with a clearly dominant dosha was misclassified by either rater. Both raters, working independently, also recorded Dwandvaja and Sannipataja labels for a substantial minority of patients, categories the manual's single-dosha rule cannot produce.

4. Discussion

DASK measures reliably. The total score achieved ICC(2,1) = 0.952 with the whole CI in the excellent band, all three subscales fell in that band, every one of the 26 items reached substantial or almost-perfect agreement, and 97.6% of the variance in the total score was genuine between-patient variance against 0.01% attributable to the rater. This is a strong result for a newly constructed instrument, and the tight CIs suggest it is not an artefact of a small or fortunate sample. The pattern of residual weakness is also familiar from that literature: of the three weakest items, Shosha and Shyava both require a graded judgement of degree on a continuous visual property, which is the item type such indices reproduce least well. Shvaitya is a different case and is addressed below.

On this evidence DASK can be used to compare groups, and it can also be used to track individual patients against a stated threshold of 8 points. MDC_{95} is a quantity rather than a verdict: it is adequate or inadequate only relative to an effect size. A treatment effect of 10 points can be resolved in individuals; an effect of 5 points cannot, and increasing the sample size improves only the precision of group means. ICC(2,k) for the total was 0.976, so averaging two independent raters would reduce MDC_{95} by roughly $\sqrt{2}$, to about 5.1 points, which may be worth the cost where individual-level precision matters most. A clinician reporting that a patient has improved by four points is, on these figures, reporting noise.

The categorical outputs warrant more caution than the scores. Both are discontinuous functions of continuous quantities: the severity band applies fixed cut-points at 35 and 70, and the dosha attribution takes the maximum of three percentages. A patient whose true total is 34.7 and one whose true total is 35.3 differ by six-tenths of a point in reality and by an entire category on the form. With SEM at 2.60 points, the classification of any patient lying within a few points of a boundary approaches a coin toss however well the underlying score is measured, and every one of the six severity disagreements occurred at a boundary. The dosha rule is the more vulnerable of the two, because it takes an argmax over three quantities that are frequently close together in this population and returns a winner regardless of margin: a patient at Vata 38% and Kapha 37% is classified Vataja with exactly the same apparent confidence as one at Vata 60% and Kapha 20%. That both raters independently reached for Dwandvaja and Sannipataja categories the algorithm cannot produce suggests that experienced clinicians recognise this and work around it. Reporting the percentage profile alongside any label, declaring a single dominant dosha only when the leading margin exceeds a threshold referenced to the measurement error reported here, and providing explicitly for dual and tripartite involvement below that threshold would raise measured classification reliability without altering a single item.

The Kapha subscale separates two properties that are routinely conflated. Its items are a heterogeneous collection: a thermometer-based temperature judgement, several tactile judgements of surface quality, a visual pallor judgement, a symptom report, a sensation of heaviness, and a judgement of temporal stability covering six months of history. They have little in common

beyond their classical attribution to Kapha, and there is no particular reason why a patient with a cold lesion should also have a sticky one. Adding eight numbers that do not covary produces a total of unclear meaning even though each component is accurately measured, and a percentage derived from such a set is a less meaningful quantity to compare against Vata and Pitta. A factor analysis on a substantially larger sample, rather than re-anchoring of items already well reproduced, is the appropriate next step. A related tension appears in item 25 (Sthairya), which is reverse anchored so that six months of quiescence earns four severity points while feeding both the Kapha subscale and the total score, so that a dosha characteristic contributes to a quantity read as severity.

The study's strengths are a fully crossed and complete design permitting separation of rater from residual variance, the reporting of both reliability and agreement in the units of the scale, separate estimation of the reliability of the categorical outputs, a paradox-resistant coefficient for every item, and independent re-derivation of all estimators from their defining formulae.

Several limitations should be borne in mind. Two raters is the minimum a reliability study can have, and it bounds generalisation to the rater population regardless of how many patients are examined; the trainer–trainee relationship between the raters reinforces the reading of these coefficients as an upper bound. Intra-rater reliability was not assessed, so between-rater disagreement cannot be separated from within-rater instability. The sample came from a single service and was about 72% psoriasis, so performance in lichen planus rests on 17 patients, and almost no patient reached the Severe band. The source records carry no timestamp, order-of-assessment or blinding field, so procedural independence rests on protocol assertion supported by the check reported in section 3.2 rather than on an audit trail, and GRRAS item 8 is accordingly only partially met. The 26 item-level analyses are exploratory, CIs being reported in place of multiplicity-corrected p values. Finally, the internal consistency analysis cannot distinguish a multidimensional subscale from a poorly anchored one.

Five revisions to the instrument follow, in order of priority: revise the dosha classification rule as described above; reappraise the Kapha subscale by factor analysis; review the near-floor items, particularly Shaitya and Shvaitya; review the seven items whose corrected item-total correlations fall below 0.30; and resolve whether the total score is a severity index or a dosha-attribution index. A confirmatory study should use more than two raters, add an intra-rater component, recruit across the full severity range with enrichment for lichen planus, and timestamp each rater's scoring event independently.

5. Conclusion

DASK measures the severity of papulosquamous disease reliably. Its total score achieved ICC(2,1) = 0.952 (95% CI 0.92–0.97), all three dosha subscales fell in the excellent band, every one of its 26 items reached substantial or almost-perfect inter-rater agreement, and no systematic rater effect was detectable in the total score. Measurement error is small enough to support individual monitoring against a threshold of 8 points (MDC_{95} = 7.21 on an integer scale). Agreement on the two categorical outputs was lower than on the continuous scores from which they are computed, each applying a hard threshold to a quantity measured with error, and the Kapha subscale was reproduced consistently between raters without cohering internally.

Tables

Table 1. DASK subscale structure and item mapping

Subscale	Constituent items	No. of items	Score range
Vata	1 Rauksha, 2 Shosha, 3 Samkocana, 4 Ayama, 5 Shyava, 6 Aruna, 13 Parushya, 14 Khara Bhava, 19 Harsha, 20 Toda, 21 Shula	11	0–44
Pitta	7 Raga, 8 Parisrava, 9 Paka, 10 Angapatana, 11 Kleda-1, 22 Daha, 23 Visra Gandha	7	0–28
Kapha	12 Shvaitya, 15 Shaitya, 16 Utsedha, 17 Sneha, 18 Kleda-2, 24 Kandu, 25 Sthairya, 26 Gaurava	8	0–32
Total	**All 26 items**	**26**	**0–104**

Items 1–12 belong to Darshana Pariksha, 13–18 to Sparshana Pariksha and 19–26 to Prashna Pariksha.

Table 2. Distribution of scores by rater (n = 60)

Scale (range)	Rater 1, mean (SD)	Rater 1, range	Rater 2, mean (SD)	Rater 2, range
Vata (0–44)	20.15 (5.96)	8–34	20.00 (5.99)	7–35
Pitta (0–28)	7.08 (4.62)	1–23	7.27 (4.95)	1–23
Kapha (0–32)	11.73 (2.74)	4–19	11.42 (2.69)	4–18
Total (0–104)	**38.97 (11.72)**	**18–74**	**38.68 (12.06)**	**19–76**

SD, standard deviation.

Table 3. Reliability, variance components and measurement error

Scale	ICC(2,1)	95% CI	ICC(3,1)	ICC(2,k)	Patient variance	Rater variance	Residual	SEM	MDC_{95}
Vata	0.945	0.91–0.97	0.945	0.972	94.5%	0.0%*	5.5%	1.40	3.88
Pitta	0.920	0.87–0.95	0.920	0.959	92.0%	0.0%*	8.0%	1.35	3.75
Kapha	0.920	0.87–0.95	0.925	0.958	92.0%	0.6%	7.4%	0.77	2.13
Total	**0.952**	**0.92–0.97**	**0.952**	**0.976**	**95.2%**	**0.0%***	**4.8%**	**2.60**	**7.21**

ICC, intraclass correlation coefficient; CI, confidence interval; SEM, standard error of measurement; MDC_{95}, minimal detectable change at the 95% level. ICC(2,1) is two-way random effects, absolute agreement, single measurement. Variance percentages are of total variance and reconcile with ICC(2,1), which is by definition the patient variance share. *The estimated rater variance component was negative on these scales and was truncated to zero.

Table 4. Bland–Altman analysis of agreement

Scale	Bias (R1 − R2)	95% CI of bias	95% limits of agreement	p (paired t)	p (Wilcoxon)
Vata	+0.15	−0.36 to 0.66	−3.75 to +4.05	0.562	0.679
Pitta	−0.18	−0.68 to 0.31	−3.94 to +3.58	0.462	0.878
Kapha	+0.32	0.05 to 0.59	−1.74 to +2.37	0.023	0.027
Total	**+0.28**	**−0.67 to 1.24**	**−6.96 to +7.53**	**0.555**	**0.163**

R1, Rater 1; R2, Rater 2. A positive bias indicates that Rater 1 scored higher. Biases are computed from unrounded item scores and may differ in the second decimal place from the difference of the rounded subscale means in Table 2. The proportional bias test was non-significant on every scale (all $p \geq 0.18$).

Table 5. Item-level agreement, all 26 items

No.	Item	Dosha	Mean R1	Mean R2	Diff.	Exact	κ_w	95% CI	AC_2
1	Rauksha	Vata	2.58	2.40	+0.18	78.3%	0.867	0.76–0.94	0.96
2	Shosha	Vata	1.12	1.03	+0.09	75.0%	0.770	0.52–0.92	0.91
3	Samkocana	Vata	2.95	2.77	+0.18	81.7%	0.887	0.80–0.95	0.98
4	Ayama	Vata	1.67	1.50	+0.17	86.7%	0.922	0.85–0.98	0.97
5	Shyava	Vata	1.72	2.05	-0.33	80.0%	0.775	0.58–0.93	0.85
6	Aruna	Vata	1.92	1.80	+0.12	73.3%	0.800	0.64–0.92	0.92
7	Raga	Pitta	1.75	1.75	0.00	75.0%	0.874	0.78–0.94	0.96
8	Parisrava	Pitta	0.85	0.87	-0.02	76.7%	0.831	0.71–0.92	0.97
9	Paka	Pitta	0.65	0.68	-0.03	88.3%	0.920	0.81–0.97	0.98
10	Angapatana	Pitta	1.07	1.10	-0.03	76.7%	0.820	0.67–0.92	0.96
11	Kleda-1	Pitta	0.85	0.93	-0.08	78.3%	0.781	0.56–0.92	0.96
12	Shvaitya	Kapha	0.28	0.32	-0.04	88.3%	0.732	0.36–0.94	0.98
13	Parushya	Vata	2.82	2.80	+0.02	78.3%	0.871	0.79–0.93	0.97
14	Khara Bhava	Vata	2.80	2.93	-0.13	73.3%	0.810	0.69–0.89	0.97
15	Shaitya	Kapha	0.10	0.07	+0.03	96.7%	0.783	0.38–1.00	1.00
16	Utsedha	Kapha	2.83	2.75	+0.08	75.0%	0.812	0.68–0.90	0.97
17	Sneha	Kapha	0.65	0.60	+0.05	88.3%	0.848	0.73–0.94	0.99
18	Kleda-2	Kapha	0.55	0.52	+0.03	96.7%	0.963	0.91–1.00	1.00
19	Harsha	Vata	0.30	0.35	-0.05	91.7%	0.824	0.67–0.96	0.99
20	Toda	Vata	1.05	1.07	-0.02	83.3%	0.850	0.69–0.94	0.98
21	Shula	Vata	1.23	1.30	-0.07	93.3%	0.962	0.92–0.99	0.99
22	Daha	Pitta	0.98	1.02	-0.04	95.0%	0.851	0.67–1.00	0.97
23	Visra Gandha	Pitta	0.93	0.92	+0.01	73.3%	0.854	0.75–0.92	0.96
24	Kandu	Kapha	2.62	2.68	-0.06	91.7%	0.927	0.82–0.99	0.98
25	Sthairya	Kapha	3.50	3.42	+0.08	91.7%	0.914	0.81–0.98	0.99
26	Gaurava	Kapha	1.20	1.07	+0.13	65.0%	0.810	0.69–0.88	0.95

Diff., mean of Rater 1 minus mean of Rater 2; Exact, percentage of patients on whom the raters recorded the identical grade; κ_w, Cohen's quadratic weighted kappa with 95% bootstrap CI (3,000 replicates); AC_2, Gwet's agreement coefficient with quadratic weights. Where AC_2 markedly exceeds κ_w the item sits near the floor and kappa is deflated.

Table 6. Internal consistency by rater

Scale	Cronbach's alpha, Rater 1 (95% CI)	Cronbach's alpha, Rater 2 (95% CI)
Vata	0.710 (0.59–0.81)	0.711 (0.59–0.81)
Pitta	0.836 (0.76–0.89)	0.810 (0.73–0.88)
Kapha	0.426 (0.18–0.62)	0.418 (0.17–0.62)
Total	**0.863 (0.81–0.91)**	**0.863 (0.81–0.91)**

Table 7. Agreement on the categorical outputs

Categorical output	Categories	Observed agreement	Kappa (95% CI)	Interpretation
Severity band (Mild/Moderate/Severe)	3	90.0%	0.800 (0.63–0.93)	Substantial
Dosha attribution (manual percentage rule)	3–4	86.7%	0.640 (0.37–0.85)	Substantial
Dosha attribution (labels as recorded)	5	85.0%	0.752 (0.58–0.89)	Substantial

McDonald's omega was 0.89 for both raters on the total scale and 0.49 and 0.51 on Kapha.
The third row uses the diagnosis labels as written on the score sheets, which admit the classical Dwandvaja and Sannipataja categories.

Figure legends

Figures 1 and 2 are supplied as separate high-resolution image files.

Figure 1. Bland–Altman plots for the three dosha subscales and the total score. The difference between raters (Rater 1 − Rater 2) is plotted against their mean. Solid line: mean difference. Dashed lines: 95% limits of agreement. Dotted line: the proportional-bias regression.

Figure 2. Item-level quadratic weighted kappa for all 26 items with 95% bootstrap confidence intervals (3,000 replicates), ordered by point estimate. Background bands correspond to the Landis and Koch descriptors.

Declarations

Ethical approval: The study was approved by the Institutional Ethics Committee of SDM College of Ayurveda Hospital and Research Centre, Udupi (approval ID [ ], dated [ ]). Written informed consent was obtained from all participants before enrolment.

Funding: [None / specify source and grant number].

Conflict of interest: [None declared / specify].

Author contributions: AKM conceived and designed the study, developed the DASK instrument and its anchor descriptors, served as Rater 1, acquired the data, performed the statistical analysis and drafted the manuscript. NS supervised the work as research guide, contributed to the study design and to the interpretation of the findings, and critically revised the manuscript for important intellectual content. S[ ] served as Rater 2 in the inter-rater observation, contributed to data acquisition and reviewed the manuscript. All authors have read and approved the final version and agree to be accountable for all aspects of the work.

Data availability: The de-identified item-level dataset and the analysis scripts supporting the findings of this study are available from the corresponding author on reasonable request.

Acknowledgments: The authors thank the attending dermatology consultant who established the reference diagnoses, the staff of the Department of Dermatology, SDM College of Ayurveda Hospital and Research Centre, Udupi, for facilitating patient access, and all patients who consented to two examinations. [Add any statistical, technical or institutional support.]

Supplementary material: Supplied as a separate file, comprising corrected item-total correlations for all 26 items (Table S1); all computed ICC forms (S2); two-way analysis of variance mean squares (S3); estimator validation details (S4); the completed 15-item GRRAS checklist (S5); and item floor and ceiling frequencies (S6).

References

1. de Vet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol. 2006;59(10):1033–9. doi:10.1016/j.jclinepi.2005.10.015

2. Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol. 2011;64(1):96–106. doi:10.1016/j.jclinepi.2010.03.002
3. Bonett DG. Sample size requirements for estimating intraclass correlations with desired precision. Stat Med. 2002;21(9):1331–5. doi:10.1002/sim.1108
4. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86(2):420–8. doi:10.1037/0033-2909.86.2.420
5. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods. 1996;1(1):30–46. doi:10.1037/1082-989X.1.1.30
6. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155–63. doi:10.1016/j.jcm.2016.02.012
7. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res. 2005;19(1):231–40. doi:10.1519/15184.1
8. de Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine: a practical guide. Cambridge: Cambridge University Press; 2011.
9. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;327(8476):307–10. doi:10.1016/S0140-6736(86)90837-8
10. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res. 1999;8(2):135–60. doi:10.1177/096228029900800204
11. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika. 1965;52(3–4):591–611. doi:10.1093/biomet/52.3-4.591
12. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968;70(4):213–20. doi:10.1037/h0026256
13. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74. doi:10.2307/2529310
14. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol. 1990;43(6):543–9. doi:10.1016/0895-4356(90)90158-L
15. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol. 2008;61(1):29–48. doi:10.1348/000711006X126600
16. Gwet KL. Handbook of inter-rater reliability. 4th ed. Gaithersburg (MD): Advanced Analytics; 2014.
17. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16(3):297–334. doi:10.1007/BF02310555
18. McDonald RP. Test theory: a unified treatment. Mahwah (NJ): Lawrence Erlbaum Associates; 1999.
19. Stuart A. A test for homogeneity of the marginal distributions in a two-way classification. Biometrika. 1955;42(3–4):412–6. doi:10.1093/biomet/42.3-4.412
20. Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60(1):34–42. doi:10.1016/j.jclinepi.2006.03.012