Published in final edited form as:

Br J Dermatol. 2008 September ; 159(4): 887–894. doi:10.1111/j.1365-2133.2008.08711.x.

Comparison of the reliability and validity of outcome instruments for cutaneous dermatomyositis

RQ Klein^{*}, CA Bangert⁺, M Costner⁺, MK Connolly[§], A Tanikawa⁰, J Okawa^{*}, M Rose^{*}, SS Fakharzadeh^{*}, D Fiorentino[¶], LA Lee, MD^{**}, RD Sontheimer⁺⁺, L Taylor^{§§}, AB Troxel, ScD^{*,§§}, and VP Werth^{*}

^{*}University of Pennsylvania School of Medicine, Philadelphia, PA, USA

⁺Department of Dermatology, University of Texas Southwestern, Dallas, TX, USA

[§]Department of Dermatology, University of California San Francisco, San Francisco, CA, USA

⁰Department of Dermatology, Keio University School of Medicine, Shinjuku-ku, Tokyo, Japan

[¶]Department of Dermatology, Stanford University, Stanford, CA, USA

^{**}Department of Dermatology, University of Colorado, Denver, CO, USA

⁺⁺Dermatology Service, Department of Medicine, Denver Health Medical Center, Denver, CO, USA

⁺⁺Department of Dermatology, University of Oklahoma, Oklahoma City, OK, USA

^{§§}Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania, Philadelphia, PA, USA

⁰Philadelphia V.A. Medical Center, Philadelphia, PA USA

Abstract

Background—Reliable and validated measures of skin disease severity are needed for cutaneous dermatomyositis (DM). CDASI (Cutaneous Dermatomyositis Disease Area and Severity Index), DSSI (Dermatomyositis Skin Severity Index), and CAT (Cutaneous Assessment Tool) skin indices have been developed as outcome instruments.

Objectives—We sought to demonstrate reliability and validity of these tools for use in measuring disease severity.

Patients/Methods—CDASI has 4 activity and 2 damage measures, with scores from 0 to 148. DSSI assesses activity based on body surface area and severity on a scale of 0 to 72. CAT uses 21 activity and damage items, for a range of 0 to 175 for activity and 0 to 33 for damage. Ten dermatologists used the instruments to score the same 12 to 16 patients in one session. Global validation measures were administered to physicians and patients.

Results—Global validation measures correlated with the 3 outcome instruments ($p < 0.0001$). CAT displayed lower inter- and intra-rater reliability relative to the CDASI. All scales correlate better with physician than patient global skin measures.

Conclusion—It appears that the CDASI may be a useful outcome measure for studies of cutaneous DM. Further testing to compare responsiveness of all three measures is necessary.

Corresponding Author: Victoria P. Werth, Hospital of the University of Pennsylvania, Department of Dermatology, 3600 Spruce Street, 2 Maloney Building, Philadelphia, PA 19104, Phone: 215-662-2737, Fax: 215-823-5171, Email: werth@mail.med.upenn.edu.

Conflicts of Interest: None

Keywords

dermatomyositis; severity index; outcome instrument; clinical trial; CDASI; DSSI; CAT

Introduction

While pathognomonic and characteristic skin findings are seen in many patients with dermatomyositis (DM), the cutaneous manifestations of autoimmune diseases are rarely studied in a systematic and methodical manner. Skin involvement may be the most severe or active component of DM and often will fail therapeutic interventions that adequately treat myositis. In addition, it has been documented that DM skin disease activity can have a significant impact on quality of life¹. Until recently, classification of DM has often focused on muscle involvement. Newer, more detailed classification systems focusing on the spectrum of cutaneous manifestations and muscle involvement have been proposed². Classic DM (CDM) refers to clinically-evident myositis based on clinical, electrophysiologic, histopathologic, and/or radiologic evaluation, while amyopathic DM (ADM) and hypomyopathic DM (HDM) refer to absent and subclinical evidence of muscle disease, respectively³. To date, it has not been possible to distinguish these subsets of patients based on hallmark cutaneous findings; however, it is unknown whether differences in pathogenesis, epidemiology, treatment, and prognosis exist by subtype of disease.

Many cutaneous disease instruments designed for inflammatory skin conditions are subjective and non-reproducible, creating discrepancies in results between different physicians, and even between the same observer⁴. Many instruments measuring skin disease are based on the model of surface-area involvement. The disadvantage of this scoring method is that many disease processes within dermatology, such as lupus erythematosus, dermatomyositis, and acne, can affect only small areas of the skin while still being significantly debilitating to patients⁵. Additionally, surface area assessments are traditionally plagued by significant inter-observer variation, making subsequent assessment of the patient by the same observer necessary⁶.

Objective observations within dermatology require careful development and validation of instruments. To evaluate potential outcome measures for the cutaneous manifestations of dermatomyositis, we assessed three cutaneous DM indices for inter- and intra-rater reliability and validity. The ultimate purpose of outcome measures is to quantitatively score a patient's skin disease as well as to provide a marker of cutaneous activity that is sensitive to changes in the patient's clinical status. An optimized outcome measure will facilitate future clinical trials and evidence-based medicine for the treatment of DM,

Materials & Methods

Ten physicians and sixteen patients were brought to the dermatology clinic at the Hospital of the University of Pennsylvania on one afternoon to complete this study. Before study commencement, physicians completed a training session with visual images of skin manifestations of DM, as well as a live patient demonstration, to familiarize themselves with the three instruments and discuss scoring methods. Physicians were given a packet divided into 16 sections (one section for each patient) with the four instruments in each division, always in the following order: CDASI, DSSI, CAT, GPhS. Physicians rotated among patient rooms, with only one physician and patient in a designated room at any one time. Patients were also given one global skin and itch score to complete once. At the end of the session, physicians were given another packet containing two more sets of instruments and the GPhS for intra-reliability purposes.

Patients

The patients were volunteers from the outpatient clinic of the department of dermatology of the University of Pennsylvania, a tertiary care center in Philadelphia, Pennsylvania. All patients had clinical presentations and histologic results from skin biopsies consistent with a diagnosis of cutaneous DM. The sixteen patients represented a full spectrum of DM skin disease activity and damage, with patients comprising all three subtypes (CDM, ADM, HDM) of DM. One patient had juvenile-onset DM, while the other fifteen patients had adult-onset disease. Fifteen of the patients were female and one was male. One patient was African-American, one was Hispanic, and fourteen were Caucasian. None of the patients included in the study had a diagnosis of another autoimmune or mixed connective tissue disease.

Physicians

All physicians were board-certified dermatologists with extensive experience treating and managing complex DM patients. Physicians had copies of the physician training session presentation and study instruments before the study session to familiarize themselves with the tools. Physician questions regarding the instruments were addressed in a group setting by the principal investigator before study commencement. All physicians scored at least twelve patients. Nine of the ten physicians re-scored two patients. At the end of the study session, physicians were given an 8-question exit survey to determine physician preference for the 3 instruments as well as the GPhS. The survey included 4 multiple choice questions and 4 short-answer questions. An informal debriefing was also held at the end of the study session, where physicians expressed their thoughts on the different instruments and study coordinators recorded the comments.

CDASI

The CDASI has separate measurements for disease activity and damage, leading to two scores for each patient. The CDASI total score (sum of activity and damage scores) captures overall disease state, while the activity score is expected to reflect current inflammatory state of disease. The CDASI describes the extent of disease in terms of the severity of involvement, but does not record the percentage of body surface area or the number of lesions, as these methods have been shown to be hard to reproduce and unreliable with diseases involving lesions of various sizes⁵. The CDASI has 4 activity (erythema, scale, excoriation, ulceration) and 2 damage (poikiloderma, calcinosis) measures for 15 anatomical locations, with total score ranging from 0 to 148, damage subscore ranging from 0 to 32, and activity subscore ranging from 0 to 116. A higher score indicates worse disease. It also specifically measures involvement of three areas – hands (Gottron's papules), periungual, and alopecia – with different measurements than the other anatomical locations.

DSSI

The DSSI assesses disease activity based on involved body surface area and severity. Body area is divided into four parts (head, trunk, upper extremity, lower extremity) and scored by percent of involvement. Severity of involvement is captured at the four anatomic locations with three symptom scores (redness, induration, scaliness). The DSSI is calculated based on the percent body surface area involved, yielding a score on a scale of 0 to 72, with higher scores representing worse disease.

CAT

The CAT uses 21 items, divided into 10 activity, 4 damage, and 7 combined lesions, to capture the characteristic DM lesions by subclass (pathognomonic lesions, erythematous lesions, vasculopathic lesions, characteristic acral lesions, etc). Grading of lesions is based on the absence or presence of a given lesion, presence of a primary finding without secondary changes,

and presence of a primary lesion associated with different degrees of secondary changes. Activity is assessed with a specific set of criteria for each lesion, including intensity of erythema or hyperpigmentation in a dark-skinned individual, presence of secondary changes, and evidence of resolution. Damage is assessed by presence or absence of atrophy and/or hyper/hypopigmentation, as well as by assessing for chronic manifestations of disease (calcinosis, lipodystrophy, atrophy, scar, poikiloderma)^{10,11}. Once again, the CAT activity and damage scores were calculated to assess overall disease state at time of assessment. Since all items that describe a particular lesion can be included for each subtype of lesion, the range of possible CAT activity and damage subscales is from 0 to 175 and 0 to 33, respectively, with higher scores indicating worse disease. The CAT also includes a visual analog scale to capture overall activity and damage; however, this is not calculated in the total CAT score and was not incorporated into this assessment, as similar global tools utilizing a visual scale were included in the overall analysis of the three instruments.

Validation Measures

A Global Physician Score (GPhS), Global Patient Score (GPaS), and Global Itch Score (GIS) were utilized to capture the overall disease activity state at the time of the study by the physicians and patients. All scores are visual analog scales ranging from 0 to 10 that have been used in previous clinical trials of rheumatologic diseases to detect small changes in disease activity^{12,13}. A score of ten represented “perfect health” for the GPhS and GPaS, while 10 represented “worst itch” for the GIS. Patients and physicians were given instructions on completing these scores and assisted when necessary. Each patient filled out the GPaS and GIS scores once, while each physician completed the GPhS while completing the other three instruments.

Assessment of Inter- and Intra- Rater Reliability

Inter-rater reliability was assessed by the group of ten physicians who scored 12 to 16 patients in one session. Of the total ten physicians, 6 scored all 16 patients, 2 scored 15 patients, 1 scored 14 patients, and 1 scored 12 patients (Table 6). All physicians individually recorded the time they spent in each patient’s room, as well as the start time for each instrument.

To assess intra-rater reliability, nine of the ten physicians scored two patients twice. The physicians arbitrarily decided which patients to rescore based on patient availability when initial physician scoring was complete. Of the 16 patients, three were not re-rated by any of the physicians, nine were re-rated by at least one physician, 3 were re-rated by two physicians, and one was re-rated by 3 physicians. In order to minimize recall of the initial scoring of patients, the physicians were not told at the beginning of the study that two additional patients had to be scored a second time at the end of the session to assess intra-rater reliability.

Assessment of Validity

Validity was assessed by comparing global physician and patient scores to each instrument as a means of identifying whether the tools were clinically and accurately reflecting current disease states. Since the global scores used in this study are widely used references, they were used to test construct validity. Construct validity seeks to measure how well one variable (or set of variables) predicts an outcome, and thus, the global measures were used to test whether the three outcome instruments were accurately reflecting what they were theoretically constructed to measure. The global measures were also used to assess whether the scoring of the outcome instruments were skewed in one direction, as this ultimately limits the capture of longitudinal disease trends.

Statistical Methods and Hypothesis

Scale Distribution

Nonparametric and parametric summary statistics along with the Shapiro-Wilk's test for normality were used to describe the sample's distribution of scores for each instrument. The Shapiro-Wilk's test is used to determine whether a sample comes from a normal distribution and is conducted by regressing the quantiles of observed data against that of the best-fitting normal distribution.

Reliability and Validity

Inter-rater reliability was assessed using the intra-class correlation coefficient (ICC) and Pearson correlations. Based on previous research, an ICC of 0.5-0.7 is considered minimally acceptable, while an ICC above 0.81 is considered almost perfect¹⁴. To analyze and describe the change in physician scores from first to second rating, test-retest intra-rater reliability was assessed using the ICC, along with correlations, crosstab analysis, and paired t-tests. In our study, ICC scores equal to or above 0.81 were considered excellent, between 0.7 and 0.81 good, between 0.5 and 0.7 moderate, and less than 0.5 were considered poor. Validity was assessed by employing Spearman's rho to correlate the CDASI, CAT, and DSSI sub- and total scores with the three validation criteria (GPhS, GPaS, GIS). An additional validation analysis was performed by grouping GPhS scores into three overall health categories – worst (0 to 3), fair (4 to 7), and good (8 to 10) – to better reflect overall disease states in a clinical setting. In this trichotomized analysis, ANOVA was used to assess linear trends and overall differences across groups for the CDASI, CAT, and DSSI total scores.

Time for Instrument Completion

To compare the length of time to complete the CDASI, CAT and DSSI, paired sample Wilcoxon signed rank tests were employed.

Results

Distribution of Scores

CDASI total score was normally distributed with total scores ranging from 0-76 (mean 25.5 ± 15.9). CDASI activity score had a normal distribution, while the CDASI damage score was abnormally distributed to the low-end of the scale. The interquartile range (25% to 75% of scores) for the CDASI total score was 15.0 to 37.5 with 90% of scores less than 49.0. The interquartile range of the CDASI activity score (possible range 0-116, study range 0-62 with mean 22.24 ± 14.2) was 13.0 to 32.5 with 90% of scores less than 43.0. CDASI damage score (possible range 0-32, study range 0-15 with mean 3.2 ± 3.2) had an interquartile range of 1.0 to 5.0 with 90% of scores less than 8.0. CDASI poikiloderma score (possible range 0-15, study range 0-13 with mean 2.9 ± 3.0) was not normally distributed and positively skewed.

DSSI was not normally distributed and skewed to the low-end of the scale with a mean score of 4.8 out of a possible score of 72. DSSI total score ranged from 0-56 (mean 4.8 ± 7.3) and the interquartile range was 0.85 to 6.15 with 90% of scores less than 11.0.

CAT activity score had a normal distribution. The interquartile range of the CAT activity (possible range 0-175, study range 0-39 with mean 13.0 ± 7.9) was 7.5 to 19.0 with 90% of scores less than 23.0. CAT damage score (possible range 0-33, study range 0-16 with mean 3.3 ± 3.4) was not normally distributed and skewed to the low end of the scale. The interquartile range was 0.0 to 5.0, with 90% of scores less than 8.0.

Three clinical validation measures – global physician score (GPhS), global patient score (GPaS), and global itch score (GIS) – were administered to physicians (GPhS) and patients (GPaS and GIS) as visual analog 0 to 10 scales. The GPhS, GPaS, and GIS demonstrated actual ranges of 0-10 (mean 6.6 ± 2.2 , with 10 representing perfect health), 0-10 (mean 4.8 ± 2.6 , with 10 representing perfect health), and 0-8 (mean 3.8 ± 2.9 , with 10 representing worst itch), respectively, on the study day.

Reliability

Inter-rater reliability (i.e. agreement among physicians) yielded an ICC in the excellent range for the CDASI, in the poor range for the DSSI, and in the moderate range for the GPhS (Table 1). For **subscore inter-rater reliability**, the ICC for CDASI activity was excellent, while CDASI damage and CAT activity were moderate, and the CAT damage was poor (Table 1). Inter-rater reliability for the term poikiloderma was poor.

Test-retest reliability studies demonstrated that the ICC for the CDASI and DSSI total scores were excellent (Table 2). For **subscore test-retest reliability**, the CDASI activity and damage ICC scores were excellent, while the CAT activity ICC was good and CAT damage ICC was moderate (Table 2). Test-retest ICC correlations indicate small variations in the physicians' test-retest scores of the same patients. Test-retest Pearson correlations ranged from $r=0.60$ to $r=0.94$, also indicating good agreement in physician test-retest scores of the same patient. The crosstab results of the test-retest reliability assessment for CDASI activity revealed 9 ratings that were either identical or within 2 points of the initial rating, with 5 scores decreasing by more than 2 points and 4 scores increasing by more than 2 points. For the CDASI damage test-retest scoring, 14 ratings were either identical or within 2 points of the initial rating, while 3 scores decreased by more than 2 points from the first rating to the second rating and one score increased by more than 2 points on the second rating. Paired t-test results revealed all but one scale with non-statistically significant differences between the physician's test-retest mean scores. The test-retest ICC for the variable poikiloderma was excellent ($r=0.81$).

Validity

Concurrent validation with associated symptoms was performed by correlating the 3 scales with the GPhS, GPaS, and GIS ($p<0.0001$ for all correlations) (Table 3). Validation of the CDASI variable poikiloderma yielded extremely poor results. All validation results were statistically significant ($p<0.0001$).

Further validation of the GPhS was performed by trichotomizing the scale into worst overall health (scores 0 to 3), fair health (scores 4 to 7), and good health (scores 8 to 10) and then testing for differences both across groups generally and assuming a linear trend for each of the 3 instruments (Table 4). CDASI activity, damage, and total scores were clearly associated with the categories; subjects with worst health had higher scores, while patients with better health had much lower scores ($p<0.0001$, linear trend <0.0001). CDASI total scores averaged 47.67 ± 14.8 for worst health, 30.49 ± 12.7 for fair health, and 14.47 ± 7.8 for good health ($r^2 = 0.53$). CAT activity and damage scores also followed this trend ($p<0.0001$, linear trend <0.0001); however, the r-squares were lower ($r^2 = 0.39$, $r^2 = 0.19$, respectively) and there was more overlap across categories. The DSSI total score had a lower r-square ($r^2 = 0.37$) and more overlap between trichotomized GPhS categories, with worst health averaging 14.87 ± 14.1 , fair health 5.44 ± 4.0 , and good health 1.28 ± 1.5 ($p<0.0001$, linear trend <0.0001).

Time for Instrument Completion

While the packets were formatted with the instruments in the same order for each patient (CDASI, DSSI, CAT, then GPhS), physicians were instructed to individually keep track of their time to complete each tool. The mean times for the CDASI, DSSI, and CAT were 6 ± 2

minutes (median=6), 2 ± 1 minutes (median=2), and 5 ± 3 minutes (median=4), respectively, for first time scoring only (Table 5). Time to complete the CDASI was significantly longer than both the DSSI and CAT times ($p<0.002$, $p<0.000$), while time to complete the CAT was significantly longer than the DSSI ($p<0.000$) (Table 5).

Physician Exit Survey

All physicians were administered an exit questionnaire at the end of the study session and nine of the ten physicians participated in the informal debriefing session following the study. 80% of physicians stated that the CDASI would be the easiest instrument to incorporate into their daily clinical environment, while 40% of physicians said the CAT and 60% of physicians said the DSSI were the most difficult instruments to use. All physicians felt that the CDASI would be the most appropriate instrument for grading the severity of improvement of DM symptoms over time. 80% of physicians felt that the global skin and itch scores were appropriate indices to capture overall skin health.

Discussion

While inter- and intra-rater reliability and consistency are fundamental principles in the development and use of outcome instruments, ease of administration, applicability of content, and demonstration of clinical responsiveness are critical components. We focused our analyses and discussions with study participants on both reliability and validity.

Validity refers to the interpretation of a test result for the purpose for which the test was designed. Since validity is not a property of the instrument, but rather the interpretation and inferences made from the test results, it must be established for each intended purpose⁷. In terms of content, each instrument used in this study was developed by leading academic dermatologists and rheumatologists with considerable expertise in DM. Since the ultimate purpose of these instruments is use in clinical trials and longitudinal patient assessment, it is crucial that scores increase and decrease with changing clinical and systemic disease activity and damage. To assess responsiveness of disease, the instruments must have a range of scores that will fluctuate with inflammatory states of DM. The goal of this study was to theoretically determine whether the tools clinically and accurately reflect current disease states by displaying a range of scores that could potentially translate into longitudinal trends; however, a clinical trial will ultimately be needed to assess whether the instruments are indeed responsive to fluctuating disease states.

While reliability is a prerequisite to validity, the internal construct of an instrument must be assessed to demonstrate clinical usefulness, inference, and appropriateness⁸. In terms of our validity analysis comparing the three outcome instruments to the GPhS, GPaS, and GIS, all scales correlated better with physician (GPhS) than patient (GPaS, GIS) global measures. The trichotomized GPhS validation assessment demonstrated that the CDASI total score had the highest correlation with the global score and the best spread of the 3 indices, thus illustrating clinical responsiveness and applicability (Table 4). Since the global measures were used as “gold standards” in this study for validation, it would also be appropriate to use the GPhS in clinical trials to correlate with the quantitative outcome measures being studied.

The interquartile range analysis also demonstrated that the DSSI and CAT activity scores do not have as large a spread as the CDASI activity score, thus also potentially limiting clinical applicability. While both the DSSI and CDASI demonstrated near perfect ICC results (0.93 and 0.86, respectively), it remains to be determined if both measures are responsive if used during a clinical trial. Further testing is required to confirm this possibility. While the CAT and CDASI were normally distributed, the CAT's lower test-retest intra-rater reliability (ICC=0.74 for activity and ICC=0.58 for damage) and inter-rater reliability (ICC=0.60 for

activity, ICC=0.43 for damage) demonstrate that this instrument performs less well in temporal stability, inter-rater agreement, and generalizability⁹.

More study is needed to develop the optimal method to measure disease activity and damage. The reversibility of some clinical findings (e.g., activity rather than damage) is unknown. Most notably, the term “poikiloderma” was controversial during training, which is reflected in our assessments. While intra-rater reliability for this term was consistent, inter-rater reliability and validity of “poikiloderma” were extremely poor. This study revealed significant confusion surrounding the term poikiloderma, signifying the need for a more precise definition.

Reviewing the experiences of the physicians for ease of administration and use of the indices, the CDASI was deemed the favored outcome instrument for clinical appropriateness. The placement of the CDASI first in the packet for each patient may have artificially decreased the time to complete the other instruments, since patients were often examined during the CDASI time period and then the two other instruments were completed without re-examination. Thus, the CDASI time is really the only accurate and valid time in this study as it was always performed first. In this regard, another limitation of this study is the fact that the instruments were not completed in a randomized order, which may have biased other aspects of the measures as well. For example, if the CDASI was performed during the examination, the examiners may have remembered the severity of the different dimensions with greater detail when completing this first instrument than the others. Moreover, fatigue may have limited the examiners' performance on later measures and limited the reliability of the latter instruments. Additionally, only a small number of patients were examined during this study, and thus, the generalizability of our results may be limited. Regardless, in combination with the normal distribution of the CDASI, the excellent ICC score, and the validity assessment, it appears that the CDASI is a useful outcome measure for studies of cutaneous dermatomyositis. Further studies are needed to determine if modifications might be beneficial and to determine responsiveness to change for all of the instruments.

Acknowledgments

Jeffrey Callen for his comments about the study design, Joseph Jorizzo for authorizing use of the DSSI, The Juvenile Dermatomyositis Disease Activity Collaborative Study Group for authorizing use of the CAT, and Misha Rosenbach and Elizabeth Gaines for assistance on the day of the study. We would also like to acknowledge the contributions of Elizabeth Dugan, who participated in the design and performance of the study.

Ethics: The protocol for this study was approved by the institutional review board of the University of Pennsylvania School of Medicine and is in accordance with the Declaration of Helsinki in its current form. All patients gave written consent before inclusion in the study.

Funding Source: This study was supported in part by grants from the National Institutes of Health (NIH K24-AR 02207) and a Veterans Affairs Merit Review Grant (Dr. Werth). Contributions to the preparation of this work were supported by The Richard and Adeline Fleischaker Chair in Dermatology Research at the University of Oklahoma Health Sciences Center (Dr. Sontheimer).

References

1. Hundley JL, Carroll CL, Lang W, et al. Cutaneous symptoms of dermatomyositis significantly impact patients' quality of life. *J Am Acad Dermatol* 2006;54:217–20. [PubMed: 16443050]
2. Sontheimer RD. Dermatomyositis: an overview of recent progress with emphasis on dermatologic aspects. *Dermatol Clin* 2002;20:387–408. [PubMed: 12170874]
3. Sontheimer RD. Cutaneous features of classic dermatomyositis and amyopathic dermatomyositis. *Curr Opin Rheumatol* 1999;11:475–82. [PubMed: 10551671]
4. Williams HC. Is a simple generic index of dermatologic disease severity an attainable goal? *Arch Dermatol* 1997;133:1451–1452. [PubMed: 9371031]

5. Albrecht J, Taylor L, Berlin JA, et al. The CLASI (Cutaneous Lupus Erythematosus Disease Area and Severity Index): An Outcome Instrument for Cutaneous Lupus Erythematosus. *Journal of Investigative Dermatology* 2005;125:889–894. [PubMed: 16297185]
6. Bhor U, Pande S. Scoring systems in dermatology. *Indian J Dermatol Venereol Leprol* 2006;72:315–321. [PubMed: 16880586]
7. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med* 2006;119:166.e7–166.e16. [PubMed: 16443422]
8. Downing SM. Reliability: on the reproducibility of assessment data. *Med Educ* 2004;38:1006–1012. [PubMed: 15327684]
9. Beckman TJ, Ghosh AK, Cook DA, et al. How reliable are assessments of clinical teaching? A review of the published instruments. *J Gen Intern Med* 2004;19:971. [PubMed: 15333063]
10. Huber AM, Dugan EM, Lachenbruch PA, et al. The Cutaneous Assessment Tool (CAT): Development and Reliability in Juvenile Idiopathic Inflammatory Myopathy. *Rheumatology*. in press.
11. Huber AM, Dugan EM, Lachenbruch PA, et al. Preliminary Validation and Clinical Meaning of the Cutaneous Assessment Tool (CAT) in Juvenile Dermatomyositis. *Arthritis Care Res*. in press.
12. Corzillius M, Fortin P, Stucki G. Responsiveness and sensitivity to change of SLE disease activity measures. *Lupus* 1999;8:655–659. [PubMed: 10568903]
13. Fortin PR, Abrahamowicz M, Danoff D. Small changes in outpatients lupus activity are better detected by clinical instruments than by laboratory test. *J Rheum* 1995;22:2078–2083. [PubMed: 8596148]
14. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–174. [PubMed: 843571]

Table 1

Inter-Rater Reliability using the Intra-Class Coefficient (ICC)

	ICC [95%CI]
CDASI Activity Score	0.84 [0.75 - 0.95]
CDASI Damage Score	0.53 [0.32-0.73]
CDASI Total Score	0.83 [0.72-0.94]
CAT Activity Score	0.60 [0.40-0.79]
CAT Damage Score	0.43 [0.22-0.64]
DSSI Total Score	0.44 [0.23-0.65]
CDASI Poikiloderma Total Score	0.49 [0.29-0.70]
Global Physician Score [0=worst to 10=perfect health]	0.61 [0.42-0.80]

ICC = Intra-Class Coefficient

95% CI = 95% Confidence Interval

Table 2

Test-Retest Intra-Rater Reliability with 18 paired test-retest ratings

	Range	M	sd	[cv]	Time1-Time2 Paired t; p-value	r	ICC [95%CI]
CDASI Activity [Scale 0-116]	3-42	20.6	11.7	[56.7]			
	Time 1						
	Time 2	18.1	10.1	[55.9]	1.84; p<0.08	0.86 [0.73-0.99]	0.84 [0.70-0.98]
CDASI Damage [Scale 0-32]	-9 to 18	2.6	5.9				
	Time 1						
	Time 2	4.3	3.9	[91.0]	0.73; p<.48	0.87 [0.74-0.99]	0.86 [0.75-0.98]
CDASI Total [Scale 0-148]	-4 to 5	0.3	1.9				
	Time 1						
	Time 2	3.9	3.5	[88.0]			
CDASI Total [Scale 0-148]	3-51	24.9	12.8	[51.5]			
	Time 1						
	Time 2	22.0	11.4	[51.9]	2.07; p<.05	0.89 [0.77-1.0]	0.86 [0.74-0.98]
CAT Activity [Scale 0-175]	-9 to 18	2.9	5.9				
	Time 1						
	Time 2	2.9	5.9				
CAT Activity [Scale 0-175]	5-25	12.3	5.3	[42.3]	1.70; p<.11	0.78 [0.61-0.94]	0.74 [0.50-0.95]
	Time 1						
	Time 2	11.0	4.4	[39.7]			
CAT Damage [Scale 0-33]	-5 to 7	1.3	3.3				
	Time 1						
	Time 2	1.3	3.3				
CAT Damage [Scale 0-33]	0-14	4.4	4.2	[96.1]	1.4; p<0.17	0.63 [0.28-0.99]	0.58 [0.27-0.89]
	Time 1						
	Time 2	3.3	3.0	[92.3]			
CAT Damage [Scale 0-33]	-2 to 11	1.1	3.3				
	Time 1						
	Time 2	1.1	3.3				

	Range	M	sd	[cv]	Time1-Time2 Paired t; p- value	r	ICC [95%CI]
DSSI Total [Scale 0-72]							
Time 1	0 - 12.1	3.8	3.6	[96.8]		0.94	0.93
Time 2	0.3-11.0	3.4	3.3	[97.5]	1.33; p<0.20	[0.88-1.0]	[0.87-0.99]
Difference	2.1 to 3.4	0.4	1.2				
CDASI Poikiloderma [Scale 0-15]							
Time 1	0-9	3.7	3.1	[83.7]		0.81	0.81
Time 2	0-9	3.4	2.8	[81.7]	0.65; p<0.52	[0.64-0.99]	[0.66- 0.97]
Difference	-4 to 4	0.3	1.8				
GPhS [Scale 0-10]							
Time 1	2-9	5.7	2.2	[38.3]		0.71	0.72
Time 2	3-9	6.8	2.1	[30.2]	0.62; p<0.54	[0.38-1.0]	[0.49- 0.95]
Difference	-1 to 5	0.2	1.5	[682.3]			

M = Mean

sd = Standard Deviation

cv = Coefficient of Variation

Paired t = Paired t-test

r = Correlation Coefficient

ICC = Intra-Class Coefficient

95% CI = 95% Confidence Interval

Table 3

Validity Assessment (n=152; 10 raters with 16 patients)

	M	(sd)	range	GPhS* r _{sp}	GPaS* r _{sp}	GIS r _{sp}
CDASI Activity Score [#] [Scale 0-116]	22.24	(14.2)	0-62	-0.77	-0.62	0.63
CDASI Damage Score [±] [Scale 0-32]	3.2	(3.2)	0-15	-0.40	-0.03	0.23
CDASI Total Score [#] [Scale 0-148]	25.5	(15.9)	0-76	-0.76	-0.53	0.62
CAT Activity Score [#] [Scale 0-175]	13.0	(7.9)	0-39	-0.69	-0.53	0.59
CAT Damage Score [±] [Scale 0-33]	3.3	(3.4)	0-16	-0.47	-0.13	0.30
DSSI Total Score [±] [Scale 0-72]	4.8	(7.3)	0-56	-0.82	-0.49	0.61
CDASI Poikiloderma [±] Total Score [Scale 0-15]	2.9	(3.0)	0-13	-0.39	-0.06	0.24
Global Physician Score [0=worst to 10=perfect health]	6.6	(2.2)	1-10			
Global Patient Score [0=worst to 10=perfect health]	4.8	(2.6)	1-10			
Patient Itch Score [0=no itch to 10=worst itch]	3.8	(2.9)	0- 8			

* Negative correlations reflect that higher scores indicate worse disease for the three DM instruments, while lower scores indicate worse global skin health for the global physician and patient assessments

[#] Scores were normally distributed

[±] Scores were not normally distributed and were skewed to the low-end of the scale

M = Mean

sd = Standard Deviation

GPhS = Global Physician Scale

GPaS = Global Patient Scale

GIS = Global Itch Scale

r_{sp} = Spearman Rank Order Correlation Coefficient

Table 4

Validity Assessment by Trichotomized GPhS

	GPhS total Mean $\pm$ sd	GPhS worst [0-3] Mean $\pm$ sd	GPhS fair [4-7] Mean $\pm$ sd	GPhS good [8-10] Mean $\pm$ sd	F-value r^2 p-value	Linear Trend
CDASI Activity [Scale 0-116]	22.24 $\pm$ 14.2	48.81 $\pm$ 12.6	26.62 $\pm$ 11.6	12.56 $\pm$ 7.1	80.00 0.5178 <0.0001	<0.0001
CDASI Damage [Scale 0-32]	3.24 $\pm$ 3.2	5.86 $\pm$ 4.8	3.89 $\pm$ 3.4	1.91 $\pm$ 2.2	14.64 0.1642 <0.0001	<0.0001
CDASI Total [Scale 0-148]	25.49 $\pm$ 15.9	47.67 $\pm$ 14.8	30.49 $\pm$ 12.7	14.47 $\pm$ 7.8	84.03 0.5301 <0.0001	<0.0001
CAT Activity [Scale 0-175]	13.03 $\pm$ 7.9	22.52 $\pm$ 7.4	15.18 $\pm$ 6.9	8.31 $\pm$ 5.1	48.27 0.3932 <0.0001	<0.0001
CAT Damage [Scale 0-33]	3.29 $\pm$ 3.4	5.33 $\pm$ 4.4	4.38 $\pm$ 3.6	1.73 $\pm$ 2.1	17.20 0.1876 <0.0001	<0.0001
DSSI Total [Scale 0-72]	4.83 $\pm$ 7.3	14.87 $\pm$ 14.1	5.44 $\pm$ 4.0	1.28 $\pm$ 1.5	44.23 0.3725 <0.0001	<0.0001

Mean $\pm$ sd = Mean $\pm$ Standard Deviation

GPhS = Global Physician Scale

 r^2 = Pearson Correlation Coefficient

F-value = Fisher's F-value

Table 5

Paired Time Comparisons using Wilcoxon Sign Rank Test (n=152 records)

Instruments	N	Median	25%	75%	Mean; sd	Range	Wilcoxon; Pvalue
CDASI vs	136	6	4	7	5.8; 2.3	1-13	10.11; 0.000
DSSI	135	2	1	2	1.9; 0.9	1-5	
CAT vs	135	4	3	6	4.9; 2.6	1-16	9.87; 0.000
DSSI	135	2	1	2	1.9; 0.9	1-5	
CAT vs	135	4	3	6	4.9; 2.6	1-16	-3.05; 0.002
CDASI	136	6	4	7	5.8; 2.3	1-13	

Note: CDASI always completed first so other times may be shorter than in reality since they were completed after initial CDASI assessment. First rating time only.

N = Number

sd = Standard Deviation

Table 6

Methodology of Patient Rating

	Physician Rater Number	Number of Patients Rated	Freq	Number of Patients Re-Rated	First Patient Re-Rated (by unique patient #)	Second Patient Re-Rated (by unique patient #)	Order of Test-Retest	Order of Test-Retest
	1	16	18	2	1	4	13, 16	4, 18
	2	16	18	2	3	13	10, 17	7, 18
	3	15	15	0	-	-	-	-
	4	16	18	2	6	7	14, 17 or 18	15, 17 or 18
	5	15	17	2	5	12	11, 15	3, 17
	6	16	18	2	7	8	3, 18	2, 17
	7	16	18	2	16	11	5, 17	7, 18
	8	12	14	2	14	9	4, 13	8, 14
	9	14	16	2	13	3	2, 15	4, 16
	10	16	18	2	13	6	9, 16	6, 18
Total	10		170	18				

Freq = Frequency