

Received Date : 28-Oct-2014

Revised Date : 10-Apr-2015

Accepted Date : 11-Apr-2015

Article type : Original Article

A Pilot Comparison Study of Four Clinician-Rated Atopic Dermatitis Severity Scales

Authors:

CY Zhao^{1,3}, AQT Tran^{1,3}, JP Lazo-Dizon^{1,3}, J Kim^{1,3}, BS Daniel^{1,3}, SS Venugopal^{1,3}, LM Rhodes^{1,3}, MG Law², DF Murrell^{1,3}

1. Department of Dermatology, St George Hospital, Sydney, Australia,
2. Kirby Institute, Sydney, Australia
3. Faculty of Medicine University of New South Wales, Sydney, Australia

Corresponding author:

Professor Dedee F Murrell

Address: Dermatology Department, St George Hospital, Ground Floor, James Laws House,
Gray Street, Kogarah, Sydney, NSW 2217, Australia

Telephone: +61-2-9113 2543 Fax: +61-2-9113-2906

Email address: d.murrell@unsw.edu.au

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/bjd.13846

This article is protected by copyright. All rights reserved.

Funding sources: This study was supported by the Independent Learning program of UNSW medical school and Premier Dermatology Clinical Trials, based at Premier Specialists. There was no pharmaceutical industry involvement.

Conflict of interest: The authors state no relevant conflicts of interest to any contents of the paper.

What's known about this topic?

- Various severity outcome measures for atopic dermatitis (AD) are currently in existence.
- A standardised severity outcome measure is a prerequisite to interpretable and comparable evidence-based medicine.
- To date, a prospective study to compare the inter-rater and intra-rater reliability of the SCORAD, EASI, TIS and SASSAD directly together has not been published.

What the study adds?

- EASI has demonstrated the highest inter-rater and intra-rater reliability, when compared to the oSCORAD, SASSAD and TIS.
- TIS is the easiest to administer.
- There is a low correlation between the patient's quality of life and physicians' objective measure of disease.

ABSTRACT

Background: There are multiple severity outcome measures for atopic dermatitis (AD). There is a need to compare the reliability of these measures.

Objectives: To compare the inter-rater and intra-rater reliability of the objective Scoring Atopic Dermatitis (oSCORAD), Eczema Area and Severity Index (EASI), Six Area, Six Sign Atopic Dermatitis (SASSAD) and Three Item Severity index (TIS); To analyse the correlation between these outcome measures to QOL instruments: Patient-Orientated Eczema Measurement, Dermatology Life Quality Index and SkinDex-29.

Methods: Twelve AD patients attended a one day scoring exercise by 5 trained dermatology clinicians. Inter-rater and intra-rater reliability were assessed using the intra-class correlation coefficient (ICC). Correlation between clinician-rated and patient-reported measures was analysed using the Spearman's rho.

Results: Regarding inter-rater reliability, EASI and SASSAD showed good reliabilities with ICCs of 0.730 (95% CI=0.500-0.900) and 0.680 (95% CI=0.440-0.880) respectively. However the ICCs were poor for TIS and oSCORAD with 0.497 (95% CI=0.233–0.785) and 0.498 (95% CI=0.234–0.785) respectively. Separate BSA component analyses showed that the oSCORAD's BSA component have contributed to its inter-rater variations. Regarding intra-rater reliability, EASI and TIS showed excellent ICCs of 0.886 (95% CI: 0.744-0.952) and 0.820 (0.614-0.923) respectively, while SASSAD showed a good reliability with ICC of 0.720 (95% CI = 0.424-0.878). However intra-rater ICC was poor for oSCORAD with 0.446 (95%CI=0.037-0.730). Regarding correlation with patient-reported measures, only SASSAD demonstrated moderate correlation with SkinDex-29, $p=0.611$ ($p=0.035$).

Conclusions: EASI demonstrated the highest inter-rater and intra-rater reliability, supporting it as the optimal AD severity outcome measure.

INTRODUCTION

Atopic dermatitis (AD) is a chronic and common dermatologic condition with various interventional options being assessed in clinical trials. However, the evidences produced are limited by the lack of a single standardised AD severity scale. This problem has prompted the Harmonizing Outcome Measures for Eczema (HOME) initiative, a collaborative group of international clinicians, patients and researchers seeking to establish a core set of AD outcome measures.¹⁻³

A recent systematic review on the validation of AD outcome measures identified 16 different AD severity scales.⁴⁻²⁰ Five of the most commonly used ones include the SCORing Atopic Dermatitis (SCORAD), Eczema Area and Severity Index (EASI), Six Areas, Six Sites Atopic Dermatitis (SASSAD), Three Item Severity index (TIS) and the Patient-oriented Eczema Measure (POEM).¹⁴⁻¹⁹ Of these, only the SCORAD and the EASI have been evaluated satisfactorily. Although the SCORAD has been shown to be a valid, internally consistent, and responsive score, its reliability has not been adequately evaluated.^{15,21,22} Also, it mixes objective and subjective symptoms in its score which may introduce bias. Although patient-reported outcomes are equally important for clinical trials, they should be measured separately from clinician-rated objective outcomes, but can be combined for a total severity score. The proposed objective-SCORAD (oSCORAD) has been shown to be consistent. However, as with the SCORAD, its reliability has also not been adequately evaluated. Meanwhile, the EASI has been shown to be a reliable, internally consistent, valid and responsive measure.^{16,22,23} It has the advantage of being completely clinician-objective, and gives separate severity scores to different body areas. When choosing between the SCORAD and the EASI, the most recent HOME consensus preferred the EASI.²⁴ However, there has been no study directly comparing these 2 scales.

To the best of the authors' knowledge, there has not been a previous study comparing a large number of AD outcome measures together. We aim to contribute to the growing process of AD outcome measures standardisation by directly comparing the inter-rater and intra-rater reliabilities of the objective measures oSCORAD, EASI, SASSAD and TIS. We will assess their correlation with

patient-oriented measures, the Dermatology Life Quality Index/Child Dermatology Life Quality Index (DLQI/CDLQI)^{25,26}, SkinDex-29²⁷ and Patient-Orientated Eczema Measurement (POEM).¹⁸

MATERIALS AND METHODS

This was a pre-planned prospective exploratory study. Ethical approval was granted by the South Eastern Sydney Local Health District HREC (Northern Sector) (HREC/12/POWH/155). The scoring was conducted on 20th October, 2012.

CLINICIAN-RATED SCORING SCALES

oSCORAD

SCORAD was developed in 1993, focusing on 3 separate areas: extent of disease, intensity, and patients' subjective symptoms (pruritus and sleep loss). The extent of the lesions was determined by the Rule of Nines. Intensity of AD was further categorised into erythema, oedema/papulation, crusting, excoriation and lichenification and is graded from 0-3, based on a representative area of the whole body.¹⁵ The oSCORAD was proposed in 1997, removing the subjective component of the SCORAD and bringing the maximal score from 103 to 83.¹⁹

EASI

The EASI is a modified version of the Psoriasis Area and Severity Index (PASI), wherein the extent of disease and objective signs (erythema, induration, excoriation & lichenification) are assessed using different weighting factors to give a maximal composite score of 72. The extent of disease is

determined through an estimation of the total body surface percentage in 4 areas: head and neck, trunk, upper and lower limbs.²³

SASSAD

SASSAD assesses 6 signs of eczema (erythema, exudation, excoriation, dryness, cracking, and lichenification) on a scale of 0-3 in 6 areas (arms, hands, legs, feet, head and neck), hence yielding a maximal possible score of 108.¹⁷ SASSAD removes the surface area component and eliminates the risk of over- or under-estimation of surface area involved.²⁸

TIS

TIS was introduced in 1999 as a simplified SCORAD, containing an easy-to-administer 4-point scale (from 0-3) on the 3 signs (erythema, oedema and excoriation), with a maximum of 9 points.¹⁴ The TIS is scored based on a representative site.

PATIENT-REPORTED SCORING SCALES

POEM

POEM assesses symptom frequency, the pattern of remission and relapse. The questionnaire consists of 7 questions, each with a 5-point scale (0-4). The questions involve the frequency of pruritus, sleep loss, bleeding, oozing, cracking, flaking and dryness. The maximal score is 28. AD severity bands according to POEM scores have also been proposed.²⁹ There is good correlation between the POEM, CDQLI/DQLI and global assessment of illness.¹⁸

CDLQI/DLQI

DLQI and CDLQI and are the most commonly used instruments to determine the impact on quality of life of dermatologic conditions. The DLQI consists of 10 questions addressing different areas such as impact on leisure activities, school / holidays, personal relationships, sleep and treatment. Each item is scored on a scale from 0 (not relevant) to 3 (very much), with a maximum of 30 points. A cartoon version of the CDQLI is available.

SkinDex-29

SkinDex-29 assesses 3 aspects of a patient's life: emotions, functioning, and symptoms. These are graded with 5-point scales giving a maximum score of 175.²⁷

SAMPLE SIZE CALCULATION

According to the COSMIN guidelines, the sample size required for reliability analysis would be less than that of a factor analysis study.³⁰ Calculations showed that with a sample size of 11 patients and 5 assessors, an observed ICC of 0.9, 0.8 and 0.7 would be estimated with 95% confidence intervals of (0.81-0.99), (0.64-0.96) and (0.48-0.92) respectively. Thus the study would be powered to discriminate a high (0.9) from moderate (0.7) ICC, which we felt was an acceptable level of precision.

PARTICIPANTS

The AD patients were selected from the senior author's specialist dermatology practice based on a range of severities and ages, with the inclusion and exclusion criteria in Supplementary Table 1.

ASSESSORS

All 5 assessors were either dermatologists or had completed full-time dermatology fellowships for 2 years. All 5 assessors have had long-term experience with scoring AD using the EASI and SCORAD. Just prior to the session, assessors also attended a one hour training session, where they reviewed all 4 scoring systems and had their queries clarified.

ONE-DAY ATOPIC DERMATITIS SCORING EXERCISE

On the scoring day, participants were placed in individual rooms asked to complete patient-oriented questionnaires. Each assessor followed a roster to rotate around the rooms. The assessors had 12 minutes to examine each patient using the 4 scales. Each scale was printed on a differently coloured sheet and these sheets were placed in random orders for each assessor, per patient. At any one point, only one assessor was assigned to one patient. Each assessors recorded the time it took for them to complete each scale.

To assess intra-rater reliability, each assessor was asked to rescore at least 4 randomly allocated patients. The randomisation process was kept undisclosed by an uninvolved investigator. Efforts were made to ensure at least half an hour time lapse between the assessors' first and second attempt to minimise recall bias. The assessors were not allowed to discuss their scores with one another.

Feedback from the assessors was formally collected in a 2 hour group meeting at the end of the scoring session.

STATISTICAL ANALYSIS

Statistical analyses were performed using Microsoft Excel and SPSS versions 18.0 and 22.0.

Reliability was determined using one way random-effect ANOVA ICC. Inter-rater reliability was graphed using scatterplot, and intra-rater reliability using Bland-Altman plots.

To determine body surface area (BSA)'s contribution to inter-rater variations, the EASI and oSCORAD had area scores graphed using scatterplot. The BSA component and the 'total minus BSA component' of the oSCORAD had its coefficient of variation (CV) calculated. The oSCORAD was chosen as only it has a directly extractable BSA component. The null hypothesis (that BSA does not contribute to oSCORAD inter-rater variations) is that the CV for the BSA component would be smaller than the 'total minus BSA component'. A paired T-test was used to determine the significance of the difference.

The correlation between clinician-oriented and patient-oriented scales were analysed using the two-tailed Spearman's rho and scatterplots.

RESULTS

Out of the 30 AD patients approached, 14 agreed to participate. On the day, 12 patients attended; 3 adults and 9 children. All patients were diagnosed with AD, with the exception of 1 patient also having ichthyosis vulgaris in addition. Table 1 shows the participants' characteristics. Four of the 5 assessors rescored 4 patients, and one assessor rescored 5.

For feasibility, all scales were completed within 3 minutes. oSCORAD and SASSAD required 2:20 minutes, EASI required 2:10 minutes and TIS 40 seconds.

For intra-rater reliability, both EASI (ICC=0.886 [95% CI: 0.744-0.952] and TIS (ICC=0.820 [95% CI: 0.614-0.923]) demonstrated excellent reliability. SASSAD showed good reliability (ICC=0.720 [95% CI: 0.424-0.878]) whilst oSCORAD (ICC=0.446 [95% CI: 0.037-0.730]) had poor reliability. The score's Bland Altman plots are shown in Figure 1-4.

For inter-rater reliability, both EASI (ICC=0.730 [95% CI: 0.500-0.900]) and SASSAD (ICC=0.680 [95% CI: 0.440-0.880]) demonstrated good reliability, whilst oSCORAD (0.498 [95% CI: 0.234-0.785]) and TIS (ICC=0.497 [95% CI: 0.233-0.785]) showed poor inter-rater reliability. Figure 5-8 demonstrated the score's inter-rater variations.

The inter-rater CV means of the oSCORAD BSA component was 0.404, higher than that of the 'total minus BSA component', which was 0.242. A paired T-Test showed a significant difference between these CV means with $p=0.01$. This showed that the BSA component of the oSCORAD contributed to its inter-rater variations. Supplementary Figures 1-2 demonstrated the oSCORAD and EASI's inter-rater BSA component variations.

Correlation between patient-reported and clinician-rated outcome measures showed moderate correlation between SASSAD and the Skindex-29, $\rho=0.611$ ($p=0.035$) (Figure 9). The mean Skindex-29 was 58.6/120 (range: 21-95), mean DLQI/CDLQI score was 12/30 (range: 5-23) and mean POEM was 17/28 (range: 9-26). There was no correlation between the objective scoring systems and the DLQI/CDLQI and POEM.

Subjective feedback from the assessors showed that all assessors had preference of assessing AD severity in different sites, as in the EASI and SASSAD, instead of grading only a representative site. Also, all assessors felt that the TIS was fast and easy to use.

DISCUSSION

Supporting the HOME consensus statement, this study showed that the EASI had consistently the most superior inter-rater and intra-rater reliability.²⁴ Interestingly, the SASSAD also showed moderate reliability, higher than the well-accustomed oSCORAD, questioning the suitability of the oSCORAD as a routine AD outcome measure.

BSA estimation has been agreed upon as an important aspect of AD severity assessment.³ However, the method of BSA estimation varies between scores. An explanation for the oSCORAD's poor inter-rater reliability may be its utilisation of the palm size according to the "rule of nines" for an exact percentage, whereas EASI estimates the body's surface area using 7 choice bands. The complexity in separately scoring palm sizes may subject the assessor to further over- or underestimation of the affected area, especially in a child.³² Despite our assessors being familiar with BSA estimation from clinical trials and the rule of nines, CV calculations have shown that the BSA did contribute to the inter-rater variations in oSCORAD scores. On the other hand, an explanation for the SASSAD's relatively good inter-rater reliability is that it accounts for BSA by including 6 separate body regions, thus accounting for extent of involvement without having to estimate an exact area.

Disease severity in AD can vary according to the site affected. The generalisation of disease severity according to a representative site, independent of the body area affected may explain oSCORAD's poor inter-rater and intra-rater reliabilities as well as the TIS' poor intra-rater reliability. On the other hand, the EASI and SASSAD attempt to allow more accurate representation of AD severity by having the different scores for all lesions at different sites. In particular, the SASSAD gives emphasis to the severity of patients' face, hands and feet in its more holistic patient assessment. These areas are usually more prominent in irritant and allergic contact dermatitis than in atopic dermatitis alone, but are particularly important when assessing impact of disease. Also, by thorough examination of disease severity of each independently affected body part, the most and least affected areas on the patients can be elucidated, thus allowing the prioritisation of targeted treatments such as UV therapy. In future validation studies, recording of the chosen 'representative sites' from SCORAD and TIS may elicit if certain areas are more suitable as 'representative sites'.

TIS' simplistic design allowed the fastest assessment time, excellent intra-rater reliability but surprisingly low inter-rater reliability. The low inter-rater reliability makes it questionable for use in multi-centre clinical trials and meta-analysis. However, TIS would benefit general practitioners as a screening tool with its good feasibility, as supported by subjective positive feedback from the assessors. However this may not apply to real clinical practice with increased amounts of distractions and interruptions.

As with utilisation of any severity scoring system, the experience of the assessors also plays a pivotal role in the efficacy of the scoring systems in practice. Even with experienced scorers, there were also variations in their subjective assessments of the lesions' severity. This could be attributed to their previous exposure to patients with AD and experiences with administering the scoring systems. Although all of the assessors have had experience with EASI and SCORAD, discrepancies also took place, suggesting that the experience levels of the assessors could not completely account for the results obtained from the study.

There were intrinsic factors that may have confounded the accurate assessment of patients' severity. These could be attributed to patients' characteristics and comorbidities. Patients of darker skin tone presented a challenge in grading lesions' erythema, as opposed to patients with lighter skin tone. Assessments relied entirely on physicians' experiences with managing patients of similar skin tone. A similar problem was encountered in one AD patient with mild ichthyosis vulgaris, wherein it was challenging to isolate AD from ichthyosis vulgaris during the scoring sessions. By including this patient in the study, the patient characteristics involved in the study therefore simulated clinical practice instead of an artificial environment. The results for this patient showed comparable inter-rater and intra-rater reliability with the rest of the study population without skewing the results. Overall, the

use of patients with heterogeneous severity is strength of the study, which mimicks clinical practice and shows that the scoring systems can discriminate patients with heterogeneity.

Interestingly, the study also revealed low correlation between patients' perception of illness through quality of life questionnaires and dermatologists' objective assessment of disease severity. Among the quality of life instruments used in the study, only Skindex-29 showed moderate correlation with SASSAD. The poor correlation between clinical severity and patients' quality of life is a well-recognised issue, and has been shown in other chronic diseases such as pemphigus.^{25,33} It is possible that the two types of scores reflected different facets of AD. The discrepancy highlights the importance of paying equal attention to both elements and collecting both scores in holistic patient management and clinical trials. Repeating this study with a larger sample size may be required to further investigate this.

An issue raised during study feedback was the responsiveness of the combinative approach used by all of these scoring systems to obtain the score. Severity of the lesions is expressed as a mix of both the reversible disease activity and the more chronic secondary changes, possibly confounding the physician's evaluation of a patient's responsiveness to change on follow-up. When chronic signs such as lichenification have more weighting, the detection of the effect of the current management regime in certain patients may be misperceived. This could possibly be addressed if these scores were modified to assess the patient's activity and secondary damage separately. This approach has already been utilised in other dermatological conditions such as the Cutaneous Lupus Erythematosus Disease Area and Severity Index (CLASI) for lupus and the Pemphigus Disease Area Index (PDAI) for pemphigus.^{34,35}

Accepted Article

One limitation of this study was its relatively small sample size. Ideally the scoring sessions should be repeated with a greater number of participants of 15-20 to allow for more precise confidence intervals. However, based on the results gathered from the study, it is possible to obtain discriminating results regarding the inter-rater and intra-rater reliability of oSCORAD, EASI, SASSAD and TIS with this sample size. Also, a larger sample size would introduce assessor fatigue given that each assessor had to score each patient using 4 different scores, plus another 2 scores repeated for intra-rater reliability testing. Another limitation was the introduction of potential recall bias by having each assessor completing one score after another which may possibly confound the results of the later score sets.

Our study implicates further follow-up studies on the inter-rater and intra-rater reliability of these outcome measures in other countries and settings. Also, a study to assess the scoring systems for patients over a period of time may be done to assess changes in disease activity and damage. Although more time consuming, this approach would provide valuable information regarding the scoring systems' responsiveness to patients' progress.

In conclusion, EASI had the highest inter-rater and intra-rater reliability in comparison with the oSCORAD, SASSAD and TIS, hence making it a valuable tool in clinical trials. TIS was the easiest to administer and would be useful as a personal marker of patients' progress in general practice by the same GP, but not in situations where there are different doctors each time or in clinical trials.

ACKNOWLEDGEMENT

We are thankful to the Department of Dermatology, St George Hospital in providing the necessary resources for the conduct of the study, and the help provided on the day from medical students Charles Wang and Clement Loh.

REFERENCES

1. Schmitt J, Spuls P, Boers M, et al. Towards global consensus on outcome measures for atopic eczema research: results of the HOME II meeting. *Allergy* 2012;67:1111-7.
2. Schmitt J, Williams H. Harmonising Outcome Measures for Eczema (HOME). Report from the First International Consensus Meeting (HOME 1), 24 July 2010, Munich, Germany. *The British journal of dermatology* 2010;163:1166-8.
3. Chalmers JR, Schmitt J, Apfelbacher C, et al. Report from the Third International Consensus Meeting to Harmonise Core Outcome Measures for Atopic Eczema / Dermatitis Clinical Trials (HOME). *The British journal of dermatology* 2014.
4. Vourc'h-Jourdain M, Barbarot S, Taieb A, et al. Patient-oriented SCORAD: a self-assessment score in atopic dermatitis. A preliminary feasibility study. *Dermatology (Basel, Switzerland)* 2009;218:246-51.
5. Verwimp JJ, Bindels JG, Barents M, Heymans HS. Symptomatology and growth in infants with cow's milk protein intolerance using two different whey-protein hydrolysate based formulas in a Primary Health Care setting. *European journal of clinical nutrition* 1995;49 Suppl 1:S39-48.
6. Sugarman JL, Fluhr JW, Fowler AJ, Bruckner T, Diepgen TL, Williams ML. The objective severity assessment of atopic dermatitis score: an objective measure using permeability barrier function and stratum corneum hydration with computer-assisted estimates for extent of disease. *Archives of dermatology* 2003;139:1417-22.
7. Silny W, Czarnecka-Operacz M, Silny P. The new scoring system for evaluation of skin inflammation extent and severity in patients with atopic dermatitis. *Acta dermatovenerologica Croatica : ADC* 2005;13:219-24.
8. Murao Y, Mae H, Konishi T, Gao X, Tanizawa T. Semiquantitative scoring in children under treatment for atopic dermatitis. *Pediatrics international : official journal of the Japan Pediatric Society* 2004;46:330-6.
9. Housman TS, Patel MJ, Camacho F, Feldman SR, Fleischer AB, Jr., Balkrishnan R. Use of the Self-Administered Eczema Area and Severity Index by parent caregivers: results of a validation study. *The British journal of dermatology* 2002;147:1192-8.
10. Costa C, Rilliet A, Nicolet M, Saurat JH. Scoring atopic dermatitis: the simpler the better? *Acta dermato-venereologica* 1989;69:41-5.

- Accepted Article
11. Charman D, Varigos G, Horne DJ, Oberklaid F. The development of a practical and reliable assessment measure for atopic dermatitis (ADAM). *Journal of outcome measurement* 1999;3:21-34.
 12. Carel K, Bratton DL, Miyazawa N, et al. The Atopic Dermatitis Quickscore (ADQ): validation of a new parent-administered atopic dermatitis scoring tool. *Annals of allergy, asthma & immunology : official publication of the American College of Allergy, Asthma, & Immunology* 2008;101:500-7.
 13. Schmitt J, Langan S, Deckert S, et al. Assessment of clinical signs of atopic dermatitis: a systematic review and recommendation. *J Allergy Clin Immunol* 2013;132:1337-47.
 14. Wolkerstorfer A, de Waard van der Spek FB, Glazenburg EJ, Mulder PG, Oranje AP. Scoring the severity of atopic dermatitis: three item severity score as a rough system for daily practice and as a pre-screening tool for studies. *Acta Derm Venereol* 1999;79:356-9.
 15. JF S, A T. Severity scoring of atopic dermatitis: the SCORAD index. Consensus Report of the European Task Force on Atopic Dermatitis. *Dermatology* 1993;186:23-31.
 16. Barbier N, Paul C, Luger T, et al. Validation of the Eczema Area and Severity Index for atopic dermatitis in a cohort of 1550 patients from the pimecrolimus cream 1% randomized controlled clinical trials programme. *The British journal of dermatology* 2004;150:96-102.
 17. Berth-Jones J. Six area, six sign atopic dermatitis (SASSAD) severity score: a simple system for monitoring disease activity in atopic dermatitis. *The British journal of dermatology* 1996;135 Suppl 48:25-30.
 18. Charman CR, Venn AJ, Williams HC. The patient-oriented eczema measure: development and initial validation of a new tool for measuring atopic eczema severity from the patients' perspective. *Arch Dermatol* 2004;140:1513-9.
 19. Schram ME, Spuls PI, Leeflang MM, Lindeboom R, Bos JD, Schmitt J. EASI, (objective) SCORAD and POEM for atopic eczema: responsiveness and minimal clinically important difference. *Allergy* 2012;67:99-106.
 20. Bahmer FA. ADASI score: atopic dermatitis area and severity index. *Acta dermato-venereologica Supplementum* 1992;176:32-3.
 21. Kunz B, Oranje AP, Labreze L, Stalder JF, Ring J, Taieb A. Clinical validation and guidelines for the SCORAD index: consensus report of the European Task Force on Atopic Dermatitis. *Dermatology* 1997;195:10-9.

22. Rullo VE, Segato A, Kirsh A, Sole D. Severity scoring of atopic dermatitis: a comparison of two scoring systems. *Allergol Immunopathol (Madr)* 2008;36:205-11.
23. Hanifin JM, Thurston M, Omoto M, Cherill R, Tofte SJ, Graeber M. The eczema area and severity index (EASI): assessment of reliability in atopic dermatitis. EASI Evaluator Group. *Exp Dermatol* 2001;10:11-8.
24. Schmitt J, Spuls PI, Thomas KS, et al. The Harmonising Outcome Measures for Eczema (HOME) statement to assess clinical signs of atopic eczema in trials. *J Allergy Clin Immunol* 2014;134:800-7.
25. Lewis-Jones MS, Finlay AY, Dykes PJ. The Infants' Dermatitis Quality of Life Index. *The British journal of dermatology* 2001;144:104-10.
26. Finlay AY, Khan GK. Dermatology Life Quality Index (DLQI)--a simple practical measure for routine clinical use. *Clin Exp Dermatol* 1994;19:210-6.
27. Chren MM, Lasek RJ, Flocke SA, Zyzanski SJ. Improved discriminative and evaluative capability of a refined version of Skindex, a quality-of-life instrument for patients with skin diseases. *Arch Dermatol* 1997;133:1433-40.
28. Charman CR, Venn AJ, Williams HC. Reliability testing of the Six Area, Six Sign Atopic Dermatitis severity score. *The British journal of dermatology* 2002;146:1057-60.
29. Charman CR, Venn AJ, Ravenscroft JC, Williams HC. Translating Patient-Oriented Eczema Measure (POEM) scores into clinical practice by suggesting severity strata derived using anchor-based methods. *The British journal of dermatology* 2013;169:1326-32.
30. Mokkink L, Terwee, C., Patrick, D., Alonso, J., Stratford, P., Knol, D., et al. . COSMIN checklist manual. Amsterdam, Netherlands: VU University Medical Center.; 2009.
31. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. *Statistics in medicine* 1998;17:101-10.
32. Ramsay B, Lawrence CM. Measurement of involved surface area in patients with psoriasis. *The British journal of dermatology* 1991;124:565-70.
33. Tabolli S, Pagliarello C, Paradisi A, Cianchini G, Giannantoni P, Abeni D. Burden of disease during quiescent periods in patients with pemphigus. *The British journal of dermatology* 2014.

34. 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. *J Invest Dermatol* 2005;125:889-94.

35. Rosenbach M, Murrell DF, Bystryn JC, et al. Reliability and convergent validity of two outcome instruments for pemphigus. *J Invest Dermatol* 2009;129:2404-10.

Legends

Table 1: Patient demographics

Baseline characteristics		AD Patients
Number	Total	12
	Paediatric	9
	Adult	3
Age	Mean	13 (SD: ± 6.4)
	Median	11
	Paediatric mean	9
	Adult mean	25
Sex (%)	Female	9 (75%)
	Male	3 (25%)
Ethnicity (%)	Caucasian	8 (67%)
	Asian	3 (25%)
	African	1 (8%)
Patient-oriented severity according to POEM* (%)	Moderate	6 (50%)
	Severe	5 (42%)
	Very severe	1 (8%)

*POEM score severity bands: 0–2 (clear/almost clear); 3–7 (mild); 8–16 (moderate); 17–24 (severe); 25–28 (very severe)

Figure 1. Bland-Altman plots demonstrating the variation in patient scores, which is a measure of the intra-rater reliability for the EASI.

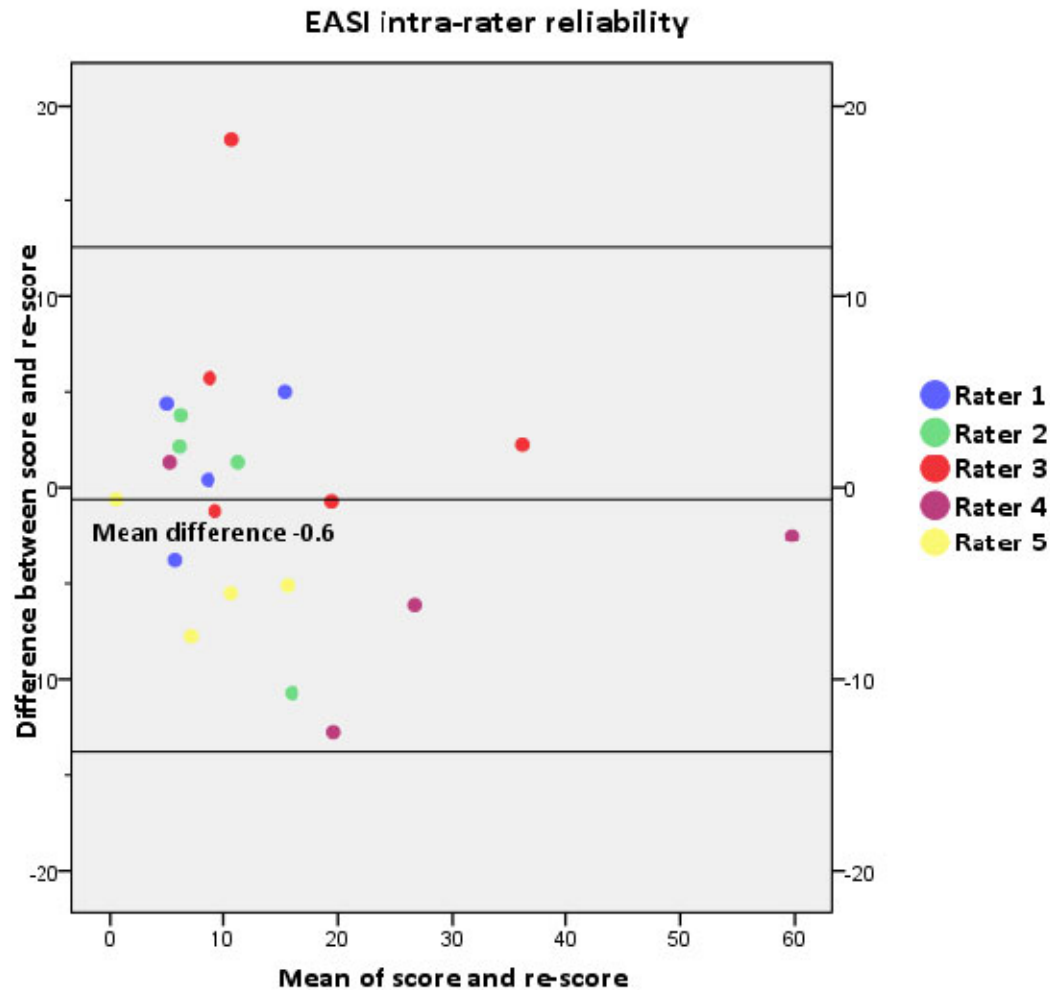


Figure 2. Bland-Altman plots demonstrating intra-rater reliability for the SASSAD.

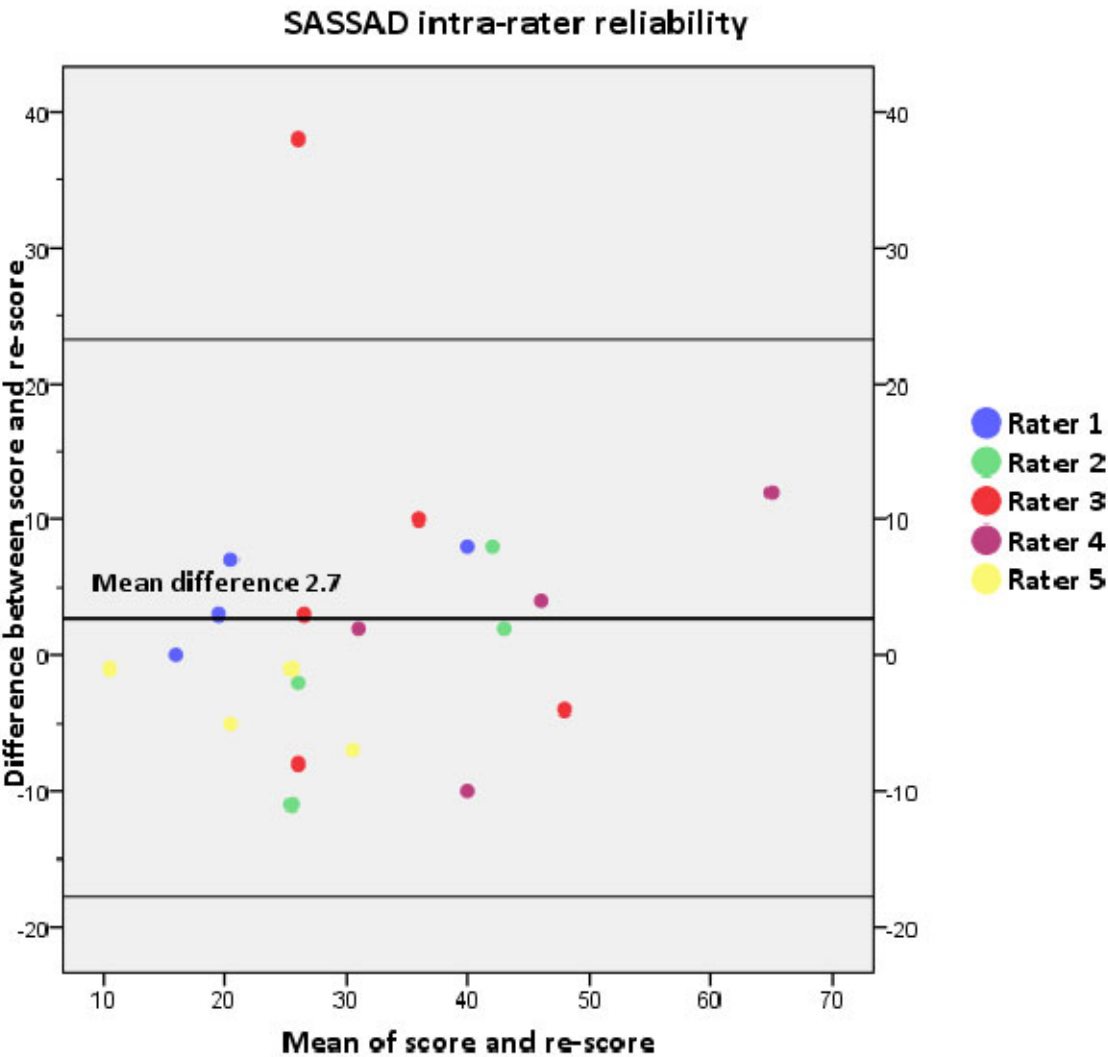


Figure 3. Bland-Altman plots demonstrating intra-rater reliability for the oSCORAD.

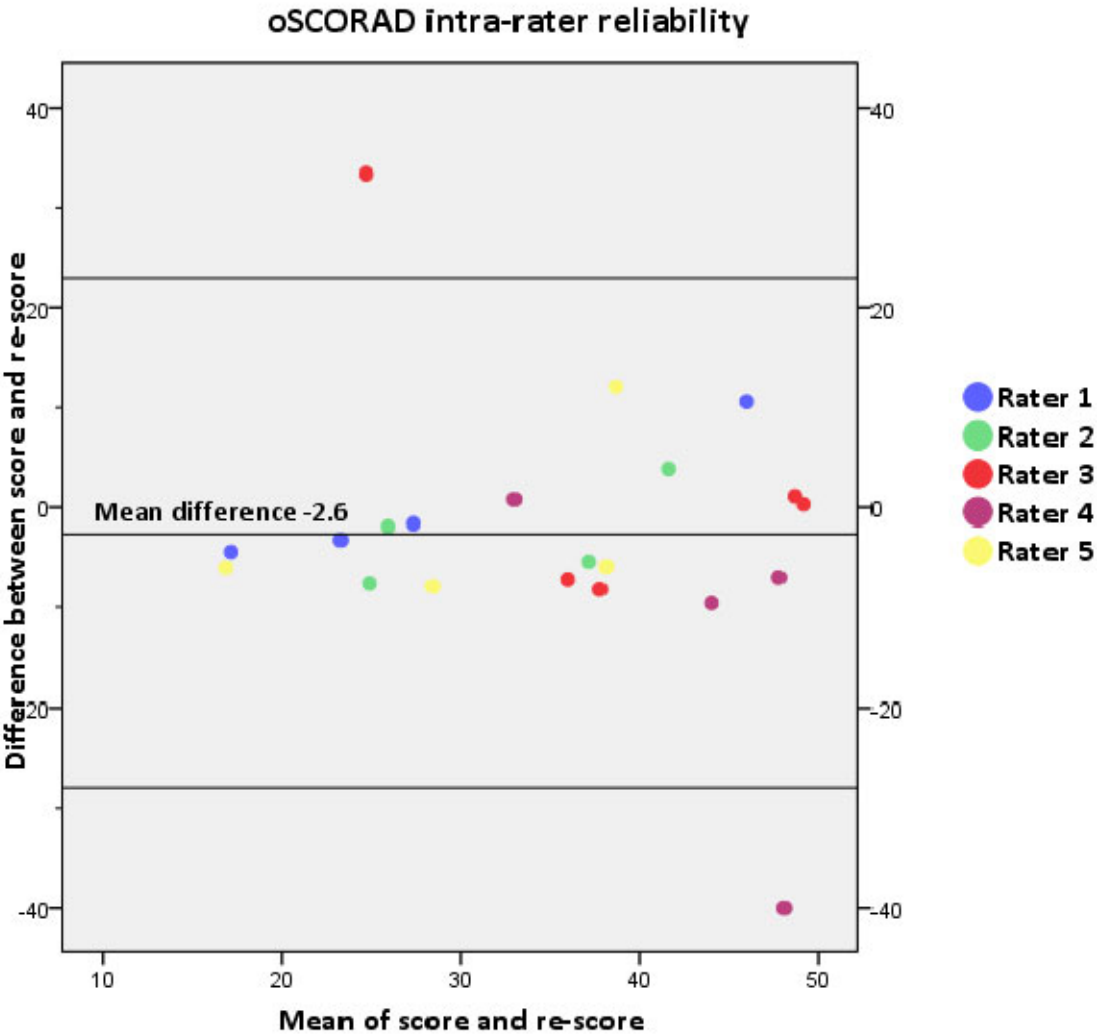


Figure 4. Bland-Altman plots demonstrating intra-rater reliability for the TIS.

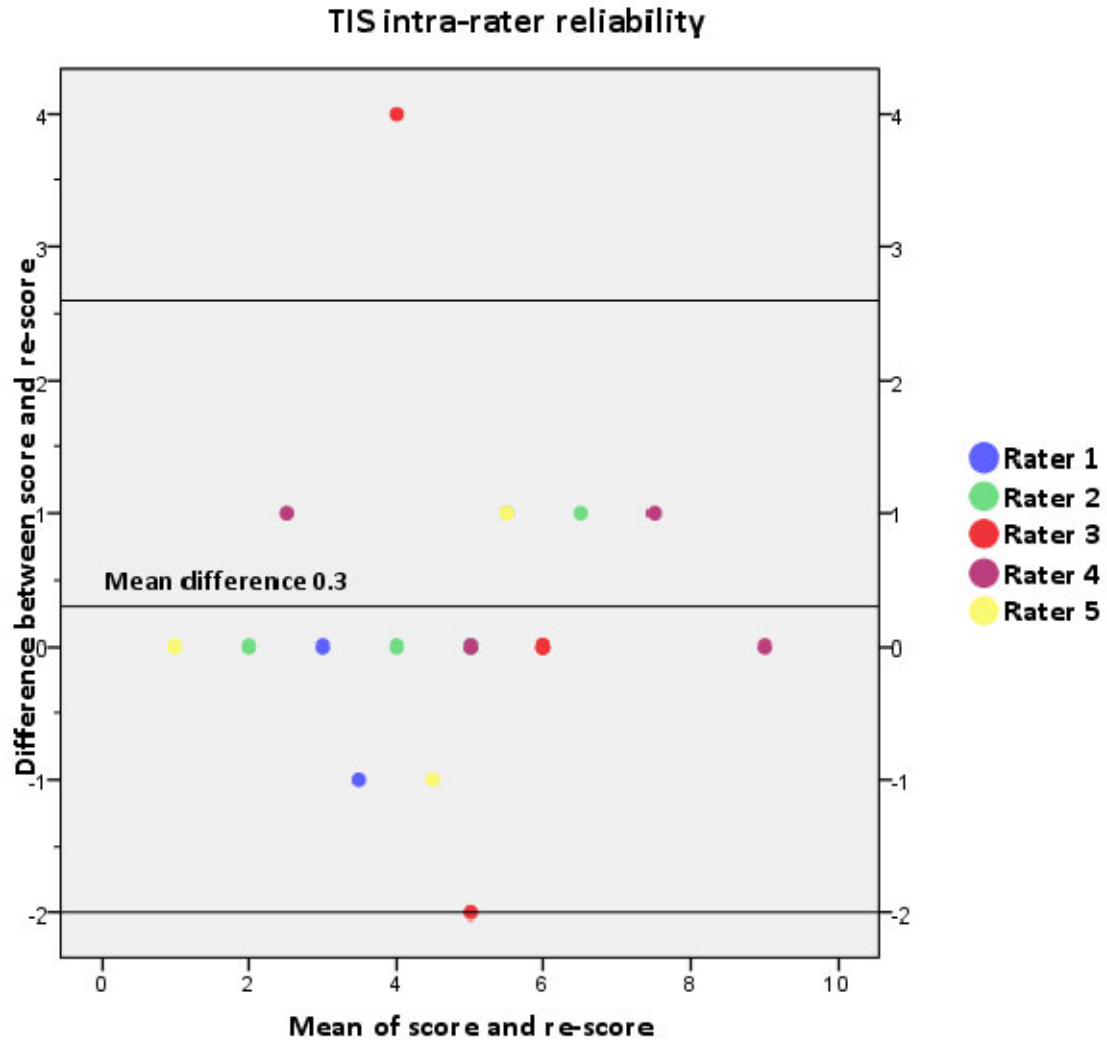


Figure 5. EASI's inter-rater scores in ascending patients' mean score rank.

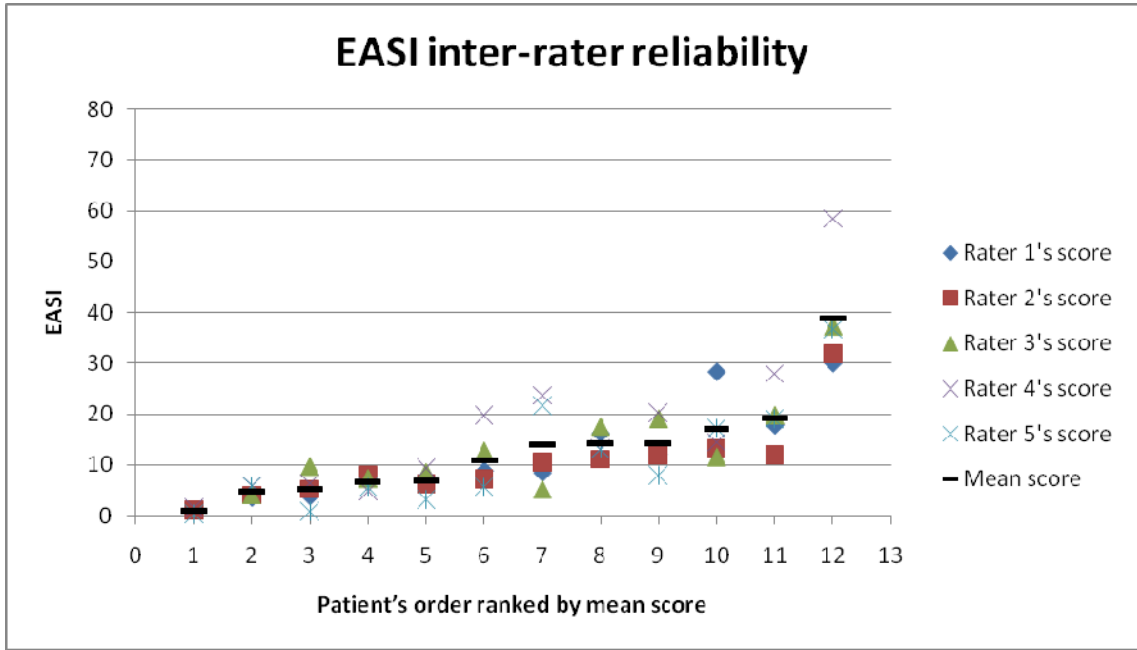


Figure 6. SASSAD’S inter-rater scores in ascending patients’ mean score rank.

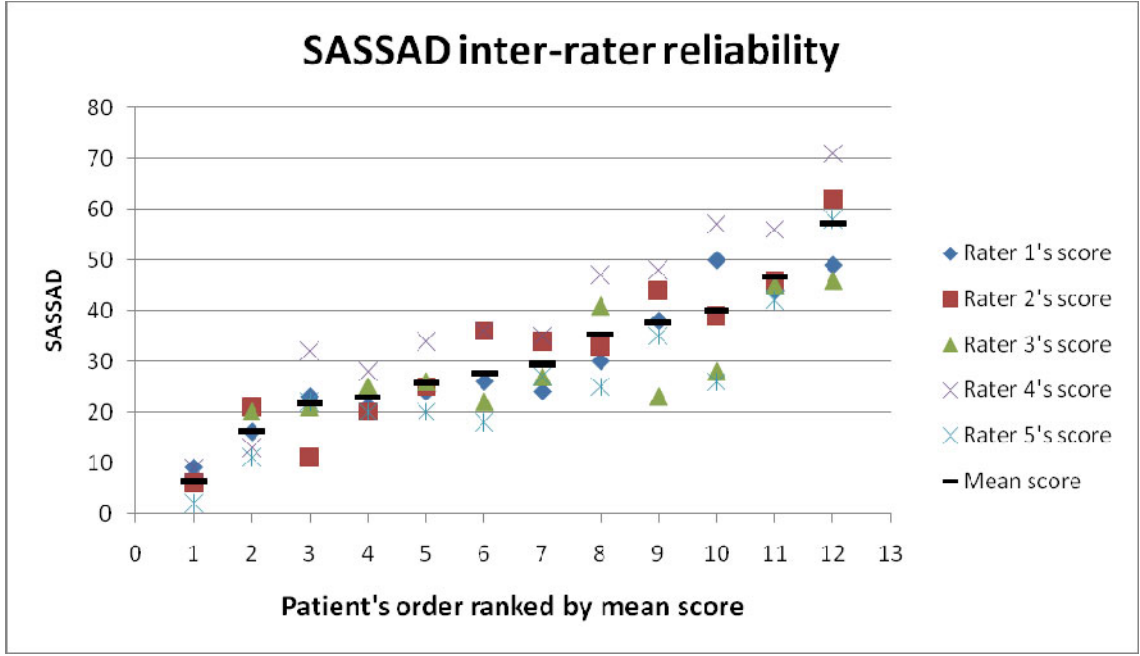


Figure 7. oSCORAD'S inter-rater scores in ascending patients' mean score rank.

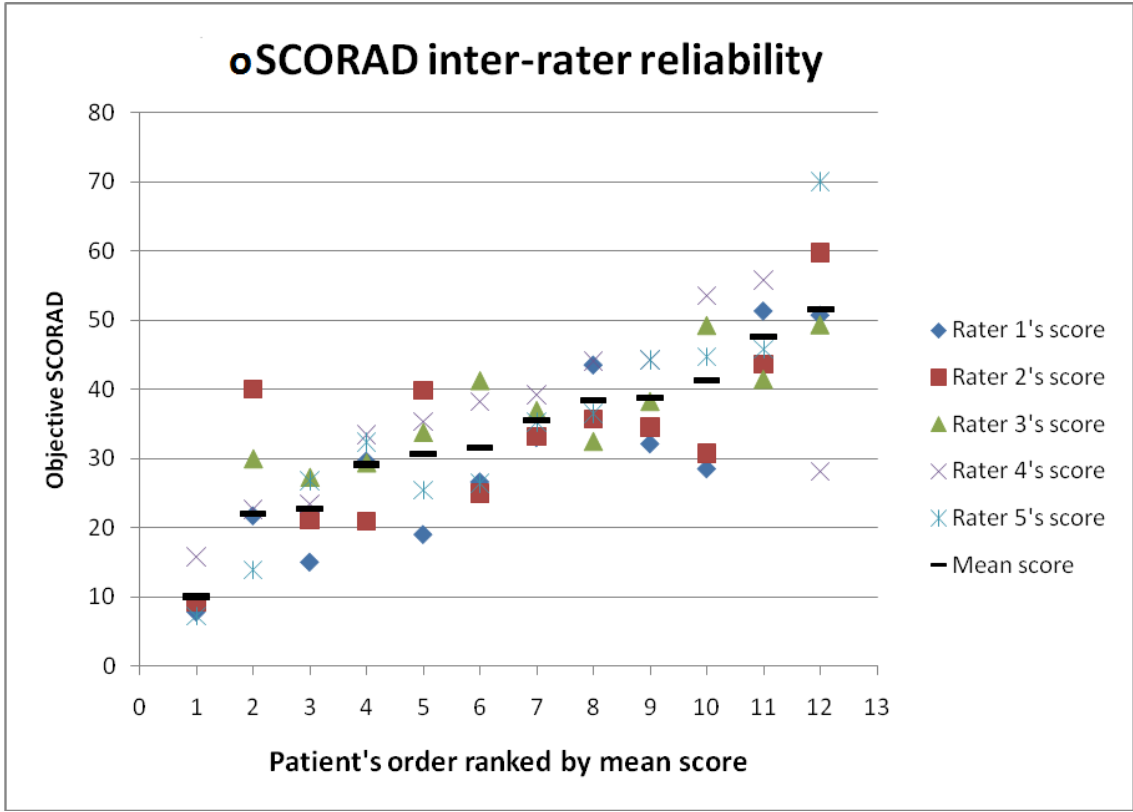


Figure 8. TIS' inter-rater scores in ascending patients' mean score rank.

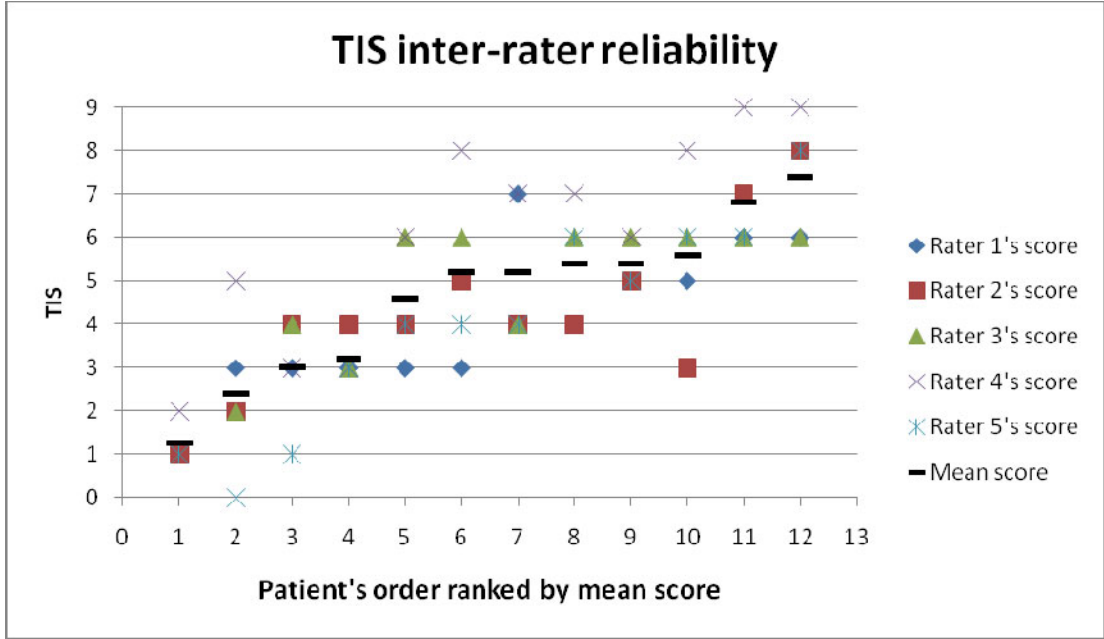
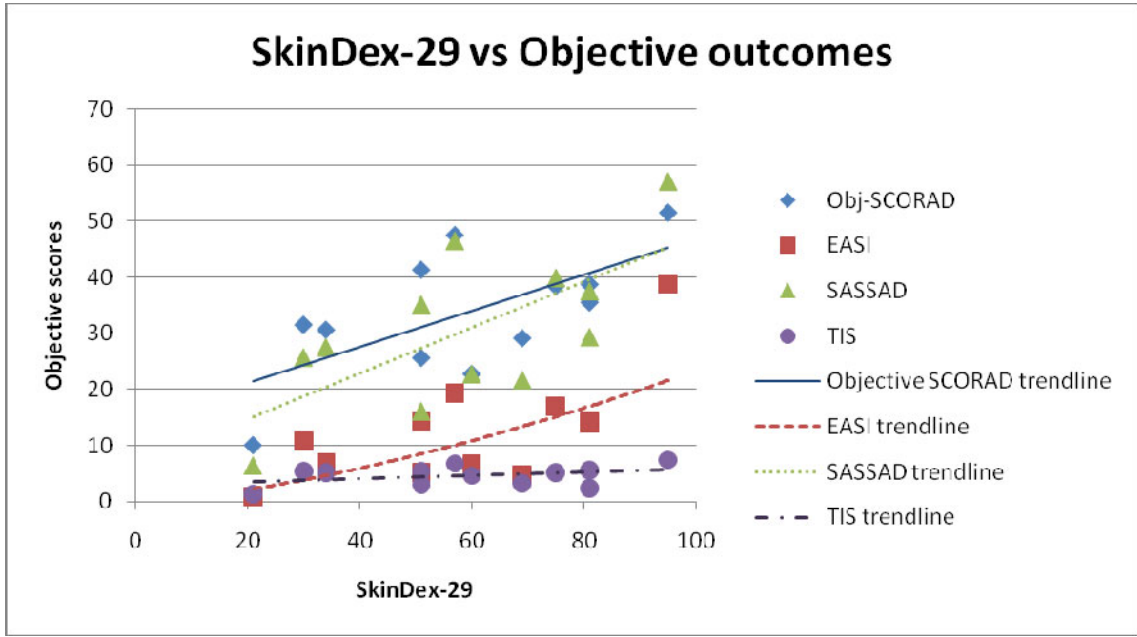


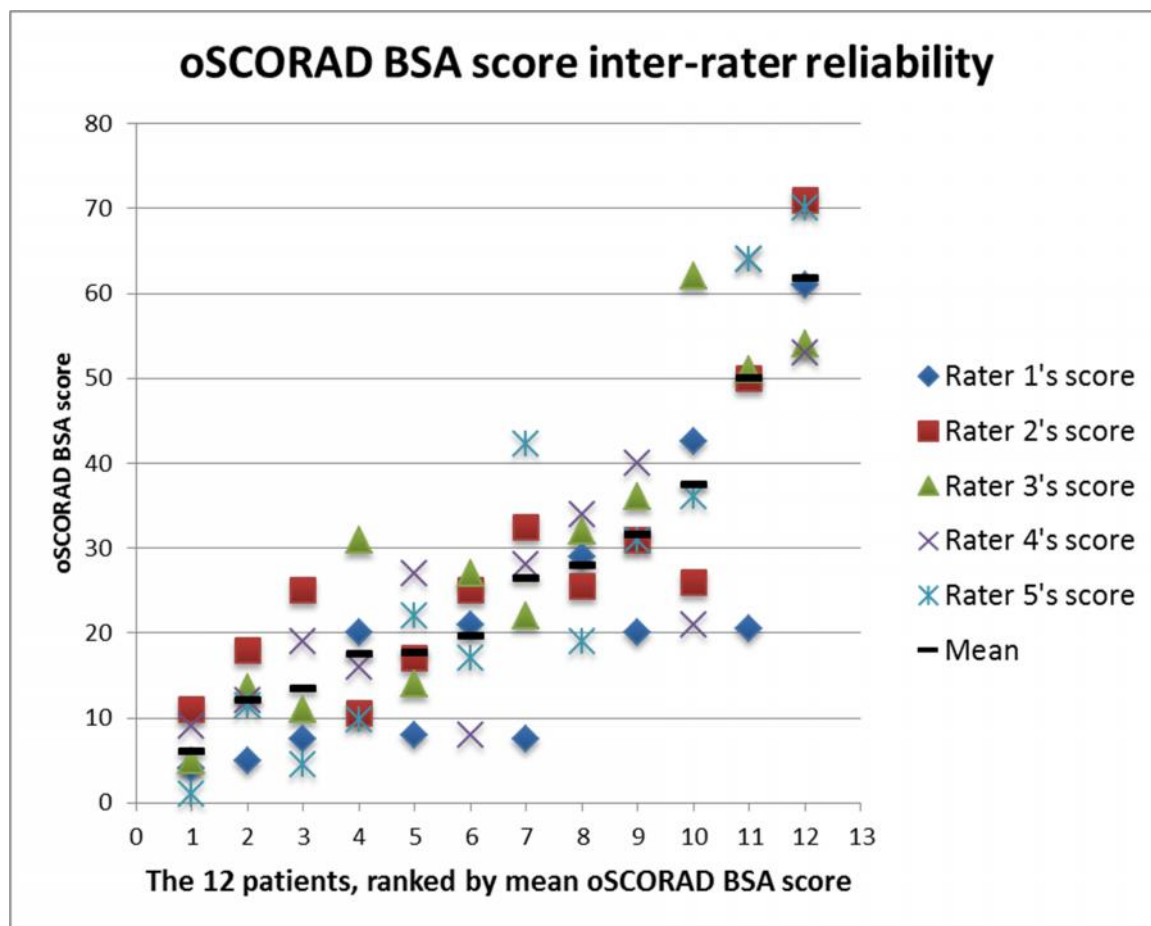
Figure 9. Correlation between Skindex-29 and objective outcomes with trend lines.



Supplementary Table 1. Inclusion and exclusion criteria for the study.

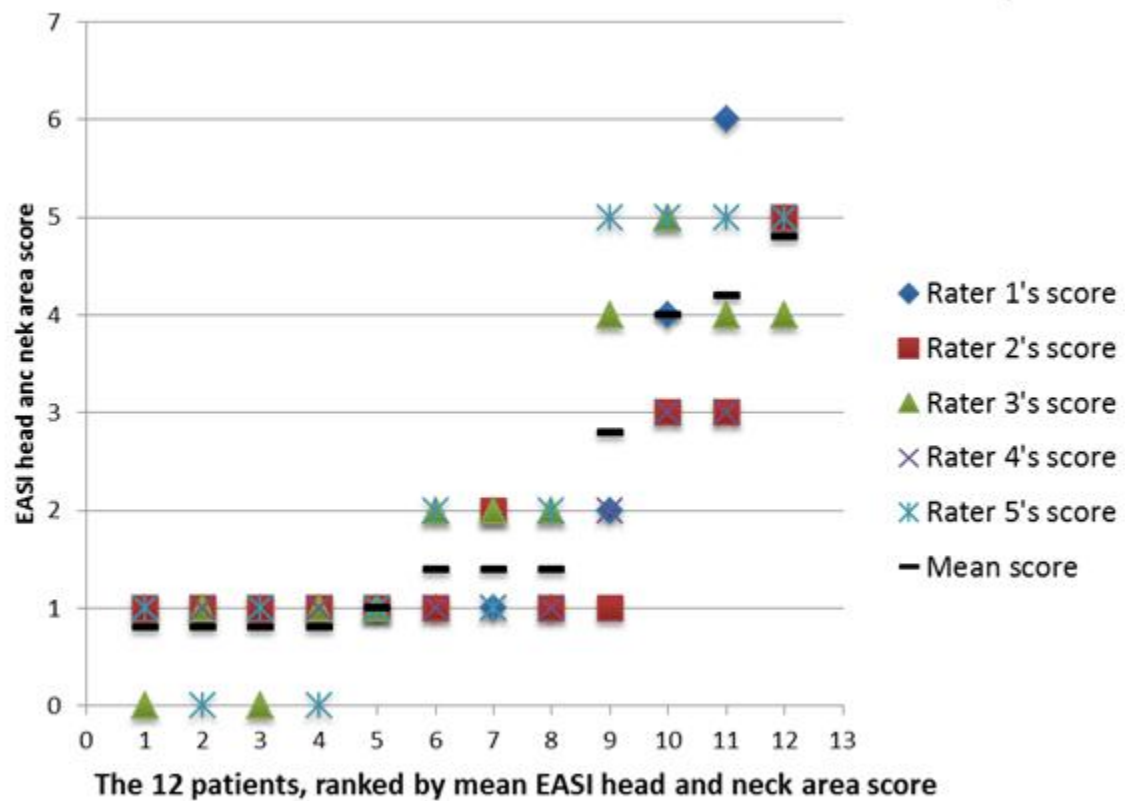
Inclusion criteria	Exclusion criteria
• Age range: 0-80 years • Those with atopic dermatitis, currently receiving or initiating treatment • Of mature mind and able to give consent, in case of children (below 18 years of age) then the guardian or caretaker can provide consent • Be able to understand and complete the DLQI/ CDLQI and Skindex-29 questionnaire (except for babies and infants)	• Participants who are unwilling or unable to comply with requirements of the protocol • Those withdrawing from the study. • Non-English speakers

Supplementary Figure 1. oSCORAD's BSA scores in ascending patients' mean score rank. The x-axes show the 12 patients ranked by their mean scores across the 5 assessors (represented by the black horizontal line sign).

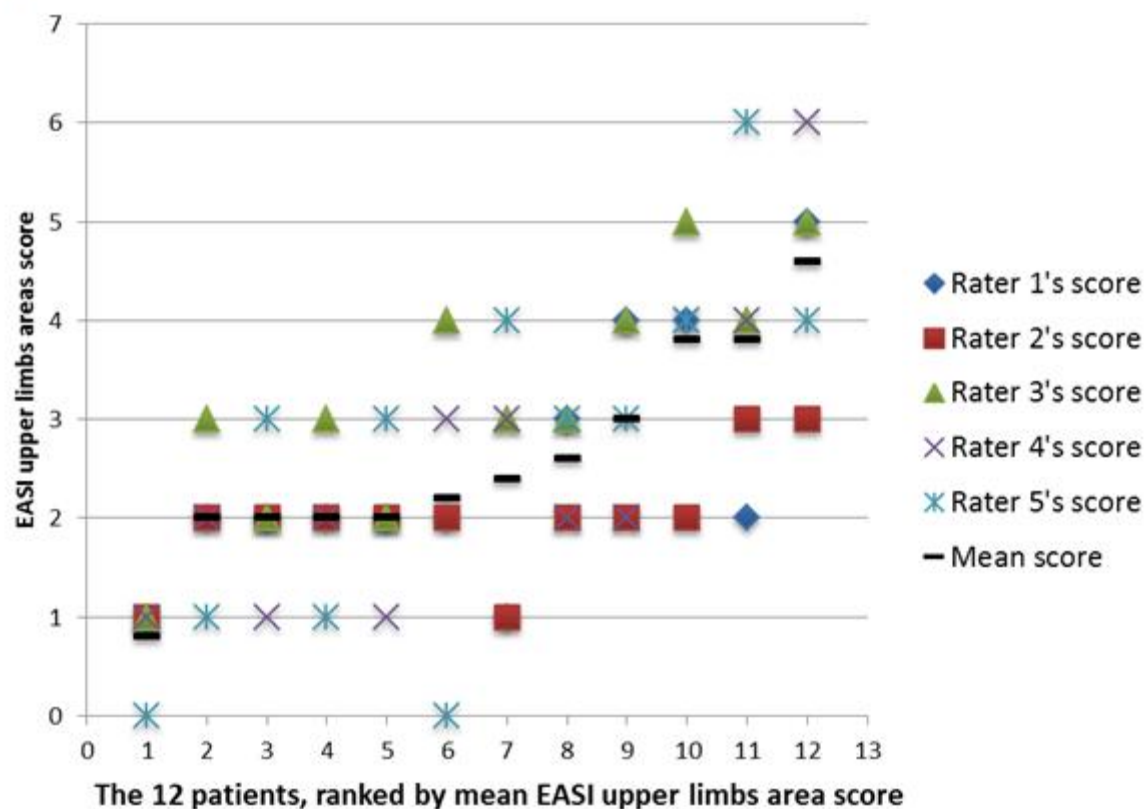


Supplementary Figure 2. EASI's area scores in ascending patients' mean score rank. The x-axes show the 12 patients ranked by their mean scores across the 5 assessors (represented by the black horizontal line sign). Each graph represents an area scored by EASI: a) head and neck, b) upper limbs, c) trunk, and d) lower limbs.

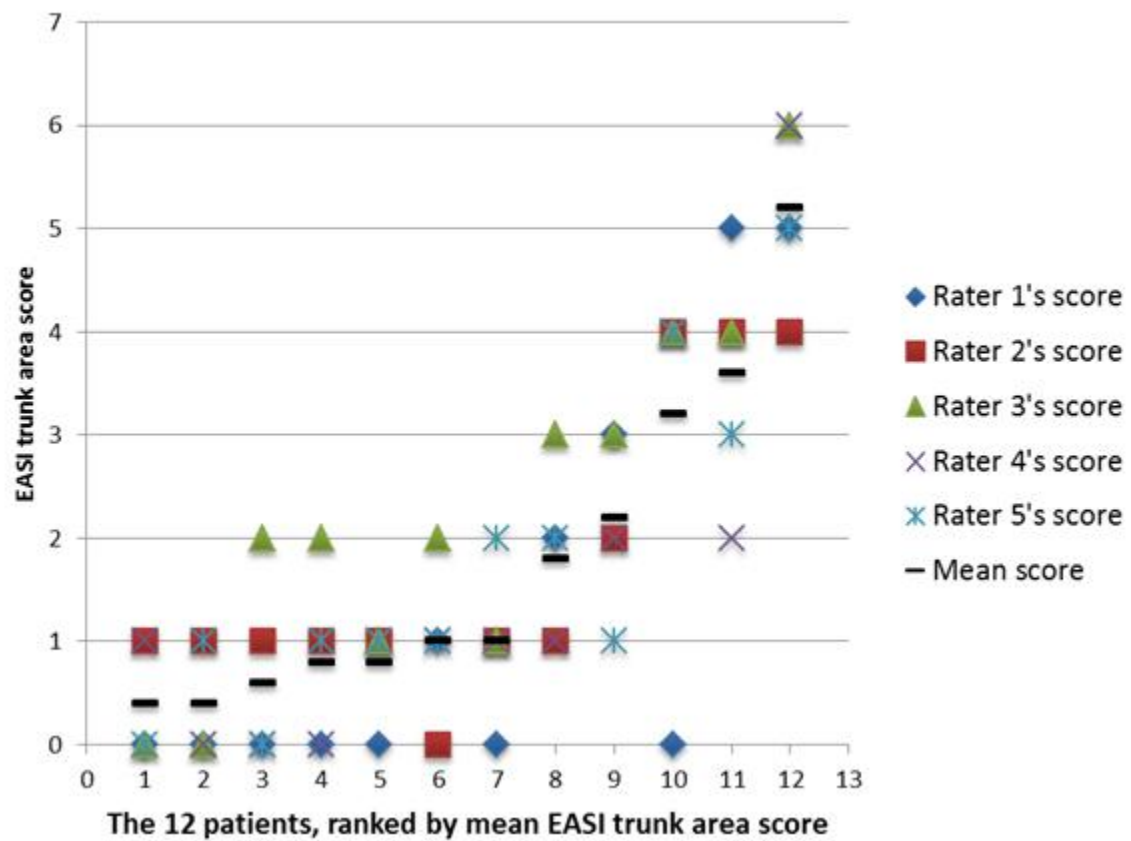
a) **EASI head and neck area score inter-rater reliability**



b) EASI upper limbs area score inter-rater reliability



c) **EASI trunk area score inter-rater reliability**



d) EASI lower limbs area score inter-rater reliability

