ORIGINAL REPORT

How Adequately is the Severity of Atopic Dermatitis Assessed by Resident Physicians Training in Dermatology? Evaluation of the Intrarater and Interrater Reliability of Clinical Scoring Systems (EASI, oSCORAD, IGA, TIS Score, SASSAD)

Katarzyna WALIGÓRA-DZIWAK[1], Katarzyna B. KUBIAK[2,3], Aleksandra DAŃCZAK-PAZDROWSKA[1] and Dorota JENEROWICZ[1]
[1]Department of Dermatology, Poznan University of Medical Sciences, Poznan, [2]Department of Public Health and Social Medicine, Medical University of Gdansk, Gdansk, and [3]Department of Computer Science and Statistics, Poznan University of Medical Sciences, Poznan, Poland

Various scoring systems for atopic dermatitis assess different aspects of the disease, but the supporting data on their validation and reliability are often limited. To ensure accurate interpretations, it is essential to conduct thorough reliability and agreement testing. This study aimed to evaluate the reliability and reproducibility of 5 atopic dermatitis assessment tools – Eczema Area and Severity Index (EASI), objective Scoring Atopic Dermatitis Index (oSCORAD), validated Investigator Global Assessment for Atopic Dermatitis (vIGA-AD), Three Item Severity (TIS) score, and Six-Area, Six-Sign Atopic Dermatitis (SASSAD) severity score – among trained dermatology residents. Thirteen patients were assessed twice by 11 raters using each scoring system. Intrarater and interrater reliability were evaluated. EASI demonstrated the highest interrater reliability (ICC = 0.768 and 0.796); SASSAD and oSCORAD followed closely (ICC = 0.741 and 0.772; ICC = 0.723 and 0.709, respectively). The TIS (ICC=0.553 and 0.584) and IGA (Kendall's W=0.273 and 0.315) exhibited the lowest reliability. Given the generally higher reliability in this study compared with previous research, implementing standardized training on atopic dermatitis scoring systems early in the education of resident physicians may improve the reliability of these tools, equipping future dermatologists to perform effectively in clinical trials and drug programmes.

Key words: atopic dermatitis; eczema; reliability; reproducibility of result.

Submitted May 26, 2025. Accepted after revision Nov 12, 2025

Published Jan 8, 2026. DOI: 10.2340/actadv.v106.43964

Acta Derm Venereol 2026; 106: adv43964.

Corr: Katarzyna Waligóra-Dziwak; Department of Dermatology, Poznan University of Medical Sciences, 60-355 Poznan, Poland. E-mail: waligora.katarzyna@spsk2.pl

SIGNIFICANCE

This study aimed to evaluate the reliability of 5 scoring tools used to assess atopic dermatitis. Thirteen patients were assessed twice by 11 dermatology residents using the Eczema Area and Severity Index (EASI), objective Scoring Atopic Dermatitis Index (oSCORAD), the validated Investigator Global Assessment for Atopic Dermatitis (vIGA-AD), Three Item Severity (TIS) score, and Six-Area, Six-Sign Atopic Dermatitis (SASSAD) severity score. The EASI had the highest reliability, followed by SASSAD, oSCORAD, TIS, and IGA. The results highlight the importance of early standardized training implemented in the education of resident physicians.

Atopic dermatitis (AD) is a chronic and recurrent inflammatory skin disease with a complex aetiology. The pathophysiology of AD stems from a multifaceted interplay of immune system dysfunction, genetic predisposition, impaired epidermal barrier function, and various environmental factors (1). It is one of the most common skin conditions worldwide, affecting approximately 2–10% of adults and up to 20% of children (2). The location and morphology of lesions vary depending on stage of the disease and patient's age. The hallmark features are chronic or relapsing course and persistent pruritus (3, 4). The other important dermatological features considered by clinical scoring systems include erythema, induration, excoriation, lichenification, xerosis, and cracking, as well as exudation (5).

No laboratory method has yet been identified for diagnosing AD. The diagnosis of the disease still relies on the clinical criteria of Hanifin and Rajka (3), formulated in 1980 and later revised following a consensus conference led by the American Academy of Dermatology, which proposed simplified criteria suitable for patients of all ages (4). The assessment of the severity of AD is based on clinical symptoms, primarily through the visual evaluation of lesions by a physician, which highlights the importance of the observer's experience. The assessment of AD severity, due to its subjectivity, may be prone to significant variability, particularly among less experienced dermatologists. As physicians training in dermatology actively participate in the treatment of patients with AD, including contribution to the implementation of drug programmes, this necessitates proper use of the primary clinical assessment tools. Evaluating the severity of the disease is a crucial component of the examination, as it helps select the appropriate therapeutic option and determines the effectiveness of treatment, both in everyday practice and in clinical studies.

To date, multiple scoring systems have been described that address different aspects of the disease, with the most popular being the EASI (Eczema Area and Severity Index) and SCORAD (Scoring Atopic Dermatitis Index) (6). Some of the scoring systems are well established and widely used for AD evaluation, while others have been applied only in single clinical protocols. Additionally, the available data on their validation and reliability as well as relative differences in disease severity scoring are often incomplete and limited (6). Specifically, the reliability of the SASSAD (Six-Area, Six-Sign Atopic Dermatitis score) and the TIS score (Three Item Severity score) has been evaluated in only a few studies, involving a limited number of observers and lacking proper investigation of intrarater and interrater reliability (5, 7).

Many scoring systems, particularly those heavily reliant on the subjective assessment of physicians, are prone to measurement errors. To prevent misinterpretations, conducting thorough reliability and agreement testing is crucial. To the best of the authors' knowledge, this is the first study examining directly together the reliability and reproducibility of AD clinical assessment tools – EASI, oSCORAD (objective SCORAD), vIGA-AD™ (the validated Investigator Global Assessment for Atopic Dermatitis, referred to later as IGA), the TIS score, SASSAD, and additionally BSA (Body Surface Area) – specifically involving resident physicians training in dermatology and following the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) (8). In the present study, we aimed to assess the reliability of 5 AD scoring systems within this rater population and evaluate relative differences in disease severity scoring across the various scales.

MATERIALS AND METHODS

The study included 13 patients (Table SI) with a confirmed diagnosis of AD and varying degrees of disease severity. The patients were approached and recruited at the Department of Dermatology at the University Clinical Hospital in Poznań. A group of 11 resident physicians training in the same Department of Dermatology were invited to take part in the study as raters on each of the 2 study days. Raters had 1–5 years of experience and were aged 28–37 years (4 men and 7 women participated on each day of the study). The number of residents participating in the study was limited by the total number of dermatology residents in the Greater Poland Voivodeship. All actively training dermatology residents who were willing to participate in the study were recruited. All physicians possessed adequate knowledge of AD dermatological manifestations. Additionally, on the first day of the study, all resident physicians underwent detailed training in the use of the analysed clinical assessment tools (EASI, oSCORAD, IGA, the TIS score, SASSAD, BSA).

Assessments were performed independently in the Dermatology Department's facilities, with no time limit set. Only 1 patient was present in the room during examination. Physicians rotated systematically between examination rooms in a predefined sequence. Over the 2 study days, 13 patients were evaluated by 11 raters using the analysed scoring systems administered in Polish on paper forms. Each patient was assessed twice on the same day by the same physician. The second assessment was performed immediately after completing the first assessment of all the patients, ensuring a time interval of at least 30 min between the ratings. The study design is presented in **Fig. 1**.

Statistical analysis

To assess intra- and interrater reliability, SPSS (version 30.0) was used (IBM Corp, Armonk, NY, USA). The intraclass correlation coefficient ICC (2,1) was calculated for quantitative data. Cohen's kappa and Kendall's W were calculated for ordinal data. The coefficient of variation (CV) was calculated to assess interrater variability.

For each score, potential differences between the 2 assessments were evaluated using a cumulative link mixed model (CLMM) with a logit link for IGA, and a linear mixed model (LMM) for EASI, oSCORAD, TIS, SASSAD. The likelihood-ratio test (LRT) was used to compare the 2 assessments.

Different scoring methods were compared using a CLMM with unequal categories and a logit link (Appendix S1). R version 4.4.1 (R Foundation for Statistical Computing, Vienna, Austria), RStudio version 2024.12.0.467, packages lme4, ordinal and emmeans were used to fit and analyse mixed models (9–13). For all analyses, a significance level of 0.05 was assumed.

A detailed description of the statistical analysis and patients' characteristics is provided in Table SI and Appendix S1. Interpretation of reliability and variability measures (14–17) is presented in the tables.

RESULTS

The mean EASI, oSCORAD, TIS, SASSAD, and median IGA scores for each patient as assessed by 11 resident physicians during 2 assessment sessions are presented in **Table I**. There were no significant differences observed between the results of the 2 assessment sessions ($p>0.05$).

The relative differences in disease severity scoring across the various scales were assessed, with reference to established scale interpretations. The EASI consistently assigned the highest disease severity ratings compared with the other scales, while oSCORAD and TIS tended to assign lower severity ratings. SASSAD assigned much lower disease severity ratings than all other scales. IGA assigned significantly lower severity than EASI, significantly higher severity than both SASSAD and TIS,

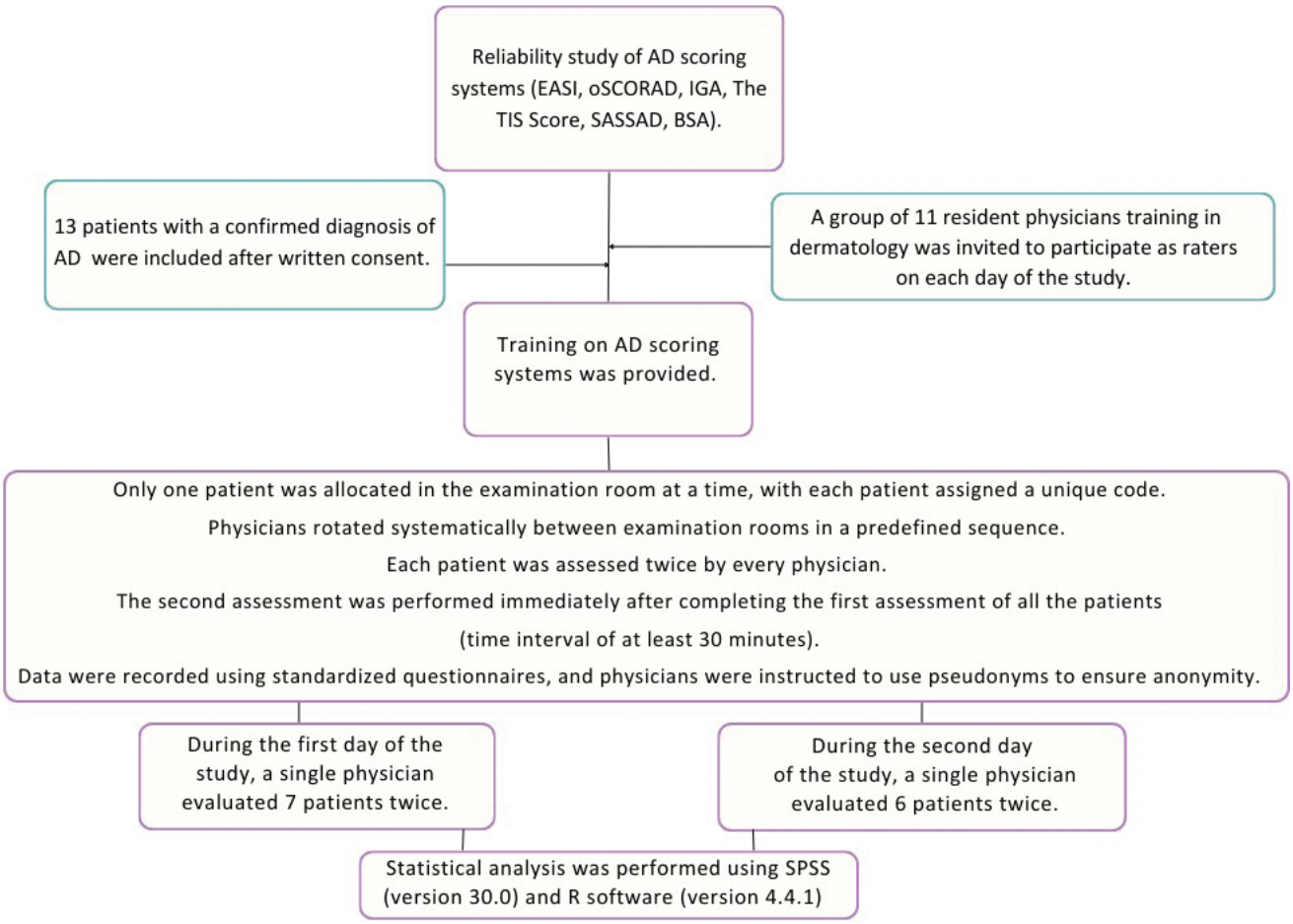

Fig. 1. Study design flowchart. AD: atopic dermatitis; EASI: Eczema Area and Severity Index; IGA: validated Investigator Global Assessment for Atopic Dermatitis; oSCORAD: Objective Scoring Atopic Dermatitis Index; SASSAD: Six-Area, Six-Sign Atopic Dermatitis score; TIS: Three Item Severity score.

and slightly higher disease severity than oSCORAD (**Table II**). There was no significant difference between the first and second assessment ($p=0.274$).

When considering interrater reliability (**Table III**), the highest ICC values were observed for EASI in both assessments (ICC=0.768 for the first assessment and 0.796 for the second), indicating very good reliability. SASSAD demonstrated similar results (ICC=0.741 and 0.772), reflecting good and very good reliability for the first and second assessments, respectively. oSCO-

Table I. Mean EASI, oSCORAD, TIS, SASSAD, and median IGA scores for each patient, as assessed by 11 resident physicians during 2 assessment sessions; *p*-values indicated.

	EASI		oSCORAD		IGA		TIS		SASSAD	
Patient no.	As.1	As.2	As.1	As.2	As.1	As.2	As.1	As.2	As.1	As.2
Patient1	26.4±10.3	27.5±8.8	40.8±11.7	44.2±9.5	3 [0.5]	3 [1]	4.5±1.6	5.0±1.7	32.3±16.8	34.0±13.0
Patient2	40.9±9.4	43.1±6.0	53.3±9.3	52.9±9.4	4 [1]	4 [1]	5.9±1.4	6.2±1.2	38.5±10.5	38.1±14.2
Patient3	5.4±5.4	6.0±4.5	16.2±8.5	20.1±10.8	1 [0]	1 [0.5]	1.6±1.0	2.3±1.7	12.5±10.0	14.9±7.4
Patient4	3.8±2.2	3.3±1.3	17.7±7.4	17.1±7.6	1 [1]	1 [1]	2.6±1.4	2.3±1.2	14.4±7.2	15.4±5.1
Patient5	17.0±7.3	14.7±7.5	31.2±7.7	30.8±9.9	2 [1]	2 [0.5]	3.1±1.4	2.8±1.4	20.5±6.1	20.4±8.2
Patient6	4.8±2.7	5.2±3.5	15.5±5.0	16.4±6.3	1 [1]	1 [1]	1.6±1.0	1.6±0.7	12.3±7.8	12.4±5.5
Patient7	7.4±3.2	8.0±4.1	21.1±8.1	21.3±8.1	1 [1]	2 [1]	2.5±1.4	2.8±1.5	17.5±7.1	18.2±8.2
Patient8	5.3±6.7	5.7±6.3	15.6±11.5	15.0±12.3	1 [0]	1 [1]	1.7±1.6	1.7±1.8	11.1±10.4	10.6±10.3
Patient9	13.8±9.6	14.7±8.8	30.6±11.2	31.0±10.9	2 [1.5]	3 [2]	2.7±1.5	3.0±1.2	18.5±11.2	20.5±6.3
Patient10	37.3±10.4	37.8±10.9	58.8±10.2	60.7±10.2	4 [1]	4 [1]	6.7±1.6	6.6±1.5	63.0±12.3	65.7±13.1
Patient11	33.3±8.5	33.6±10.8	52.6±13.3	50.8±14.4	3 [1]	4 [1]	5.5±2.2	6.0±1.8	59.4±14.8	61.5±10.6
Patient12	11.6±7.5	9.8±6.8	21.1±7.6	20.6±8.7	1 [1]	1 [1]	2.4±1.0	2.5±1.6	16.5±5.0	18.9±9.1
Patient13	8.6±3.2	7.1±3.7	26.0±10.0	24.4±8.4	2 [0.5]	2 [1]	3.0±2.0	2.5±1.7	21.5±6.7	24.5±9.0
P-value	0.929		0.675		0.267		0.419		0.117	

As.1: Assessment 1; As.2: Assessment 2; CLMM, cumulative link mixed model; EASI: Eczema Area and Severity Index; IGA: validated Investigator Global Assessment for Atopic Dermatitis; IQR: interquartile range; LMM: linear mixed model; LRT: likelihood-ratio test; oSCORAD: Objective Scoring Atopic Dermatitis Index; SASSAD: Six-Area, Six-Sign Atopic Dermatitis score; TIS: Three Item Severity score. Results presented as means±standard deviations or medians [IQR]. Indicated are P values of the LRT assessing the differences between two assessments through a mixed model (CLMM with a logit link for IGA and LMM for EASI, SASSAD, SCORAD, TIS) with disease severity as outcome variable, assessment as fixed effect, and patient and rater as random effects.

Table. II. *Post hoc* pairwise comparisons among disease severity scoring scales (EASI, oSCORAD, IGA, TIS, and SASSAD) in a CLMM with unequal categories and a logit link, with the IGA score set as the reference score

Pair	Estimate	Std. error	*p*-value	Interpretation
IGA – EASI	–3.027	0.207	<0.0001	IGA assigns lower severity than EASI
IGA – SASSAD	4.937	0.274	<0.0001	IGA assigns higher severity than SASSAD
IGA – oSCORAD	0.366	0.180	0.2516	No significant difference between IGA and oSCORAD
IGA – TIS	1.884	0.196	<0.0001	IGA assigns higher severity than TIS
EASI – SASSAD	7.965	0.335	<0.0001	EASI assigns much higher severity than SASSAD
EASI – oSCORAD	3.393	0.210	<0.0001	EASI assigns higher severity than oSCORAD
EASI – TIS	4.911	0.240	<0.0001	EASI assigns higher severity than TIS
SASSAD – oSCORAD	–4.572	0.265	<0.0001	SASSAD assigns lower severity than oSCORAD
SASSAD – TIS	–3.053	0.250	<0.0001	SASSAD assigns lower severity than TIS
oSCORAD – TIS	1.518	0.189	<0.0001	oSCORAD assigns higher severity than TIS

Results are averaged over 2 assessments. P-values were adjusted with Tukey's adjustment for multiple comparisons. The greater the absolute value of the estimate, the greater the difference between assigned disease severity by a pair of scales.
EASI: Eczema Area and Severity Index; IGA: validated Investigator Global Assessment for Atopic Dermatitis; oSCORAD: Objective Scoring Atopic Dermatitis Index; SASSAD: Six-Area, Six-Sign Atopic Dermatitis score; Std. error: standard error; TIS: Three Item Severity score; CLMM: cumulative link mixed model. The quantitative scores were transformed into ordinal data with reference to established scale interpretations: The EASI severity was categorized as 0=clear, 0.1–1.0=almost clear, 1.1–7.0=mild, 7.1–21.0=moderate, 21.1–50.0=severe, and 50.1–72.0=very severe; the oSCORAD severity was categorized as <15=mild, 15–40=moderate, and >40=severe; the IGA severity was categorized as 0=clear, 1=almost clear, 2=mild, 3=moderate, and 4=severe; the TIS score severity was categorized as <3=mild, 3–6=moderate, and >6=severe; the SASSAD severity was categorized as 0=clear, 1–36=mild, 37–72=moderate, and 73–108=severe.

RAD showed slightly lower but still good reliability (ICC=0.723 and 0.709). The lowest reliability scores, indicating moderate to poor interrater reliability, were observed for TIS (ICC=0.553 and 0.584) and IGA (Kendall's W=0.273 and 0.315). Additionally, BSA interrater reliability was assessed, showing good reliability in both assessments (ICC=0.738 and 0.712).

Although most scales demonstrated generally good interrater reliability, this does not indicate an absence of variation between raters (**Table IV**, Fig. S1). The EASI, TIS, and SASSAD exhibited high interrater variability, with coefficient of variation (CV) values ranging from 44.0 (SASSAD) to 53.8 (EASI). The IGA and oSCORAD showed moderate interrater variability, with CV values of 34.1 and 36.5, respectively. Moreover, BSA had a CV of 43.0, indicating high interrater variability.

Intrarater reliability was found to be very good across all scales (**Table V**), with high ICC or Cohen's kappa values. The EASI demonstrated the highest reliability (ICC=0.955), followed by oSCORAD (ICC=0.932), SASSAD (ICC=0.925), TIS (ICC=0.856), and IGA (Cohen's kappa=0.862). Additionally, BSA showed very good intrarater reliability (ICC=0.937).

When considering the individual dermatological score components (erythema, oedema/papulation, excoriation, lichenification, exudation, dryness, and cracking), interrater reliability was generally moderate, with some exceptions (see Table III). The highest Kendall's W values were observed for exudation (0.497 and 0.599), indicating moderate to good interrater reliability, while the lowest values were recorded for cracking (0.309 and 0.237) and excoriation (particularly in the first assessment, 0.133), indicating poor to moderate interrater reliability. Intrarater reliability was higher than interrater reliability for all score components (see Table V), with Cohen's kappa values ranging from 0.790 to 0.854. Good intrarater reliability was observed for oedema/papulation and dryness, while all other features demonstrated very good intrarater reliability.

Table III. Interrater reliability assessment for EASI, oSCORAD, IGA, TIS, SASSAD, BSA, and dermatological score components

	Reliability measure		Reliability interpretation	
Scale	As.1	As.2	As.1	As.2
	ICC			
EASI	0.768	0.796	Very good	Very good
oSCORAD	0.723	0.709	Good	Good
TIS	0.553	0.584	Moderate	Moderate
SASSAD	0.741	0.772	Good	Very good
BSA	0.738	0.712	Good	Good
	Kendall's W			
IGA	0.273	0.315	Poor	Moderate
	Reliability measure		Reliability interpretation	
Feature	As.1	As.2	As.1	As.2
	Kendall's W			
Erythema[a]	0.305	0.371	Moderate	Moderate
Oedema/papulation[a]	0.423	0.459	Moderate	Moderate
Excoriation[a]	0.133	0.404	Poor	Moderate
Lichenification[a]	0.338	0.285	Moderate	Poor
Exudation[b]	0.497	0.599	Moderate	Good
Dryness[b]	0.297	0.376	Poor	Moderate
Cracking[b]	0.309	0.237	Moderate	Poor

As.1: Assessment 1; As.2: Assessment 2; BSA: body surface area; EASI: Eczema Area and Severity Index; ICC: intraclass correlation coefficient; IGA: validated Investigator Global Assessment for Atopic Dermatitis; oSCORAD: Objective Scoring Atopic Dermatitis Index; SASSAD: Six-Area, Six-Sign Atopic Dermatitis score; TIS: Three Item Severity score. [a]EASI-derived features; [b]SASSAD-derived features. ICC interpretation: <0.40, poor reliability; 0.40–0.59, moderate reliability; 0.60–0.74, good reliability; 0.75–1.00, very good reliability. Kendall's W interpretation: < 0.3, low reliability; 0.3 to 0.5, moderate reliability; >0.5, large reliability.

Table IV. Interrater variability assessment for EASI, oSCORAD, IGA, TIS, SASSAD, and BSA

Scale	Variability measure Coefficient of variation, %	Variability interpretation
EASI	53.8±27.3	High
oSCORAD	36.5±15.9	Moderate
TIS	50.5±21.1	High
SASSAD	44.0±21.0	High
BSA	43.0±23.4	High
IGA	34.1±17.3	Moderate

BSA: body surface area; EASI: Eczema Area and Severity Index; IGA: validated Investigator Global Assessment for Atopic Dermatitis; oSCORAD: Objective Scoring Atopic Dermatitis Index; SASSAD: Six-Area, Six-Sign Atopic Dermatitis score; TIS: Three Item Severity score. Coefficient of variation interpretation: 0–20% slight variability, 21–40% moderate variability, 41–60% high variability, and above 60% very high variability.

Table V. Intrarater reliability assessment for EASI, oSCORAD, IGA, TIS, SASSAD, BSA, and dermatological score components

Scale	Reliability measure	Reliability interpretation
	ICC	
EASI	0.955±0.030	Very good
oSCORAD	0.932±0.052	Very good
TIS	0.856±0.071	Very good
SASSAD	0.925±0.036	Very good
BSA	0.937±0.038	Very good
	Cohen's kappa	
IGA	0.862±0.090	Very good
Feature	**Reliability measure**	**Reliability interpretation**
	Cohen's kappa	
Erythema[a]	0.824±0.097	Very good
Oedema/papulation[a]	0.790±0.261	Good
Excoriation[a]	0.839±0.090	Very good
Lichenification[a]	0.822±0.078	Very good
Exudation[b]	0.842±0.114	Very good
Dryness[b]	0.798±0.089	Good
Cracking[b]	0.854±0.072	Very good

BSA: body surface area; EASI: Eczema Area and Severity Index; ICC: intraclass correlation coefficient; IGA: validated Investigator Global Assessment for Atopic Dermatitis; oSCORAD: Objective Scoring Atopic Dermatitis Index; SASSAD: Six-Area, Six-Sign Atopic Dermatitis score; TIS: Three Item Severity score. [a]EASI-derived features; [b]SASSAD-derived features. ICC interpretation: <0.40, poor reliability; 0.40–0.59, moderate reliability; 0.60–0.74, good reliability 0.75–1.00, very good reliability. Cohen's kappa interpretation: 0.01–0.20, none to poor reliability; 0.21–0.40, fair reliability; 0.41–0.60, moderate reliability; 0.61–0.80, good reliability; 0.81–1.00, very good reliability.

DISCUSSION

There are various tools available for assessing AD, each designed to measure multiple aspects of the condition in distinct ways. The ideal tool for evaluating AD should strike a balance between comprehensiveness and ease of implementation, ensuring it can be used effectively in everyday clinical practice. While there is available research on AD assessment by trained dermatologists, the impact of early standardized training for physicians undergoing education in dermatology on their ability to accurately utilize these tools has yet to be explored. In the present study, we aimed to evaluate the reliability of 5 AD scoring systems within this rater population and assess relative differences in disease severity scoring across the scales.

In this study, resident physicians tended to show higher reliability for most scores than in previous studies, demonstrating good to very good intra- and interrater reliability for EASI, oSCORAD, and SASSAD. However, TIS and IGA showed poor to moderate interrater reliability, which could result from their simplistic and general approach, with TIS assessing disease severity based on a representative site and IGA relying on a broadly defined global evaluation, highlighting the need for thorough location-by-location assessment and BSA inclusion in scoring systems to ensure interrater reliability.

The high reliability of EASI in our study aligns with the 2014 HOME (Harmonizing Outcome Measures for Eczema) group recommendation that clinical trials should include at least EASI to assess AD severity, as it demonstrates adequate validity, internal consistency, intraobserver reliability, and responsiveness (18). Additionally, the HOME group reports that oSCORAD has adequate validity and interobserver reliability, but its intraobserver reliability is uncertain (18,19). In the present study, we report a high intraobserver reliability of oSCORAD.

Consistent with our findings, Zhao et al. (20) reported good interrater reliability for EASI and SASSAD (ICC=0.730 and 0.680, respectively) and low reliability for TIS (ICC=0.497) in a study with 12 AD patients and 5 trained dermatology physicians. In another analysis, the interrater reliability of TIS was generally assessed as intermediate (7) with 2 analysed studies indicating intermediate to adequate interrater reliability (21, 22). In comparison, in the same analysis, based on a study with 6 trained raters and 6 patients, the interrater reliability of the total SASSAD score was described as acceptable, while the interrater reliability of the SASSAD components was found to be inadequate (7). In terms of intrarater reliability, Zhao et al. (20) reported both EASI and TIS to show very good reliability (ICC=0.886 and 0.820, respectively), while SASSAD demonstrated good reliability (ICC=0.720), which was slightly lower than in our study. In contrast to our results, oSCORAD exhibited poor interrater and intrarater reliability (20).

Similar findings regarding EASI were reported by Hanifin et al. (23), who evaluated EASI in a study of 20 patients assessed by 15 dermatologists, showing good overall interrater reliability. In their study, induration/papulation was the most challenging to assess, with the lowest kappa values, while erythema and excoriations showed good reliability, and lichenification showed reasonable reliability (23). In contrast, our study found excoriations to have low reliability, especially in the first assessment. In fact, in our study, the interrater reliability of individual dermatological score components was generally moderate, whereas the reliability of total scores of most scales was good. This difference can be attributed to variations in how residents evaluated the intensity of specific features, with some assigning higher scores to certain features while underestimating others, ultimately resulting in total scores with good reliability. Additionally, good BSA reliability contributed to higher reliability of BSA-including scores. This suggests that while physicians maintained a general sense of severity, their approach to evaluating specific dermatological characteristics varied, highlighting the need for more intense and standardized training to enhance consistency in feature-based assessments.

Regarding the IGA, Simpson et al. (2020) tested the reliability of the scale (16), which previously lacked standardization, using an online survey based entirely on photographic documentation. The assessments were often limited to 1 body area, with a general description of disease extent. The first and second surveys showed high interrater reliability (Kendall's W: 0.809 and 0.819)

and high intrarater reliability (ICC: 0.879 and 0.881). However, the use of photographs limited raters' ability to assess disease severity and simplified the assessment, especially while providing a description of disease extent. This may explain the lower IGA interrater reliability observed in our study.

While EASI, oSCORAD, and SASSAD exhibited good to very good interrater reliability in this study, high reliability does not preclude variability between raters. Interrater reliability reflects consistency in relative rankings, yet raters may still assign differing absolute scores. Consistently assigning higher scores than other raters may maintain alignment in rankings, introducing systematic bias and contributing to overall variability. In this study, rater 11 consistently rated higher, while rater 1 assigned lower scores (see Fig. S1), contributing to interrater variability despite good reliability. Similar levels of variability were reported elsewhere (24).

Limitations

One of the limitations of this study was the relatively modest size of the study sample. However, it is noteworthy that the present study included a greater number of raters and subjects compared with several previous reliability studies of AD questionnaires, particularly those involving the SASSAD and TIS scores (7, 20, 24). The study's findings indicate that it is feasible to derive meaningful conclusions concerning the reliability of analysed scales with the sample size used. While increasing the number of participants could potentially improve statistical accuracy and reduce the likelihood of differences arising by chance, it may also contribute to increased assessor fatigue. Each physician was already required to evaluate each patient using 5 distinct scoring systems twice on each day of the study, amounting to 70 scoring assessments on the first day and 60 scoring assessments on the second day per physician. Over the 2 days, each of the 13 participating patients was assessed in total 110 times across various AD scales, yielding 1,430 assessments overall, as 11 raters performed 2 evaluations for each patient using 5 different scoring systems. The number of residents participating in the study was limited by the total number of dermatology residents in the Greater Poland region. All actively training dermatology residents willing to participate in the study were recruited.

A potential limitation was recall bias, as raters conducted both assessments on the same day. However, conducting the assessments twice in 1 day was necessary to ensure the stability of the patients' condition, given the frequent fluctuations in severity characteristic of AD. The time interval and numerous evaluations between assessments likely minimized this risk.

The selection of dermatology residents from a single department may have influenced scoring consistency due to shared training and clinical environment. However, the decision to include raters from the same department offers the advantage of enabling a more controlled evaluation of the effects of early standardized training. Future studies across multiple institutions could further assess the role of training standardization.

Another limitation of the study is the inclusion of patients of Caucasian ancestry with a Fitzpatrick skin type of I–III. Since assessing the severity of AD in individuals with higher phototypes can be more challenging, particularly in evaluating erythema (25), the results might differ if patients from a more diverse range of backgrounds were included.

Conclusion

EASI demonstrated the highest interrater and intrarater reliability, followed by oSCORAD and SASSAD. Simpler scales like TIS and IGA had lower interrater reliability but maintained strong intrarater reliability, emphasizing the need for step-by-step location-based assessment and preferably BSA inclusion in scale design. Dermatological features showed generally moderate interrater reliability, with cracking and excoriations being the most challenging to assess. Interrater variability across scoring systems ranged from moderate to high. Relative differences in disease severity scoring across various scales were observed.

The study suggests that clinical trials could benefit from assigning a single investigator per patient, as intrarater reliability appeared consistently higher than interrater reliability. Additionally, using relative improvement thresholds (e.g., EASI90, a 90% reduction from baseline) may be preferable to using absolute mean differences, as it could help mitigate inconsistencies between raters. As EASI showed the highest reliability and reproducibility, it is the recommended scale for use in clinical settings, preferably assessed by a fixed investigator per patient and interpreted in the context of relative improvement thresholds to reduce interrater variability effects. As oSCORAD demonstrated good reliability and lower interrater variability, it could be considered when assigning a single investigator is not possible. Although SASSAD generally showed good reliability, it should be used with caution as it does not account for BSA. Simpler scales, like IGA and TIS, require further investigation and may be useful for quick patient assessments in daily practice but should rather be avoided as sole assessment measures. Additionally, researchers should be aware that relative differences in disease severity scoring may occur depending on the scale used.

As reliability of analysed scores tended to be generally higher (apart from IGA) than in previous studies, after 1 session of training, implementing standardized training on AD scoring systems early in the education of resident physicians might improve the reliability of these tools,

though further studies are needed to exclude chance effects. Early expert-guided training in dermatology could help reduce inconsistencies in atopic dermatitis assessment, supporting more uniform evaluations of disease severity and better preparedness to implement new therapies.

ACKNOWLEDGEMENTS

The authors extend their gratitude to all the physicians and patients who volunteered to participate in this study. A Large Language Model (LLM) was used to enhance readability and language.

IRB approval status: The study has been conducted according to the guidelines for human research and the World Medical Association Declaration of Helsinki. The study protocol was approved by the bioethical committee of Poznan University of Medical Sciences (KB – 509/23).

Informed consent statement: All the subjects have completed an informed consent form before joining the study.

Conflict of interest disclosures: KW-D has received funding to attend scientific conferences/participated in clinical trials conducted for the following pharmaceutical companies: Leo Pharma, Sanofi, AbbVie, and GSK. KBK declares no conflict of interest. AD-P has been a speaker/received funding to attend scientific conferences/participated in clinical trials conducted for the following pharmaceutical companies: Leo Pharma, Galderma, Bausch Health, L'Oréal, Medac, Sun-Farm, GSK, Novartis, Pfizer, Eli Lilly, Astellas Pharma, AstraZeneca, Sanofi, Takeda, Eucerin, Almirall, and AbbVie. DJ has been a speaker/received funding to attend scientific conferences/participated in clinical trials conducted for the following pharmaceutical companies: Leo Pharma, Sanofi, Galderma, and GSK.

REFERENCES

1. Kim J, Kim BE, Leung DYM. Pathophysiology of atopic dermatitis: clinical implications. Allergy Asthma Proc 2019; 40: 84–92. https://doi.org/10.2500/aap.2019.40.4202
2. Raimondo A, Lembo S. Atopic dermatitis: epidemiology and clinical phenotypes. Dermatol Pract Concept 2021; e2021146. https://doi.org/10.5826/dpc.1104a146
3. Hanifin JM, Rajka G. Diagnostic features of atopic dermatitis. Acta Derm Venereol 1980; 60: 44–47. https://doi.org/10.2340/00015555924447
4. Eichenfield LF, Hanifin JM, Luger TA, Stevens SR, Pride HB. Consensus conference on pediatric atopic dermatitis. J Am Acad Dermatol 2003; 49: 1088–1095. https://doi.org/10.1016/S0190-9622(03)02539-8
5. Yang YB, Lynde CW, Fleming P. Common atopic dermatitis rating scales: a practical approach and brief review. J Cutan Med Surg 2020; 24: 399–404. https://doi.org/10.1177/1203475420923644
6. Bożek A, Reich A. Assessment of the severity of atopic dermatitis. Dermatol Rev 2016; 6: 479–485. https://doi.org/10.5114/dr.2016.63839
7. Schmitt J, Langan S, Deckert S, Svensson A, von Kobyletzki L, Thomas K, et al. Assessment of clinical signs of atopic dermatitis: a systematic review and recommendation. J Allergy Clin Immunol 2013; 132: 1337–1347. https://doi.org/10.1016/j.jaci.2013.07.008
8. 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: 96–106. https://doi.org/10.1016/j.jclinepi.2010.03.002
9. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using Lme4. J Stat Softw 2015; 67: 1–48. doi:10.18637/jss.v067.i01
10. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2024 [accessed February 7, 2025]. Available from: https://www.r-project.org/
11. Posit Team. RStudio: Integrated Development Environment for R. Boston, MA: Posit Software, PBC; 2024 [accessed February 7, 2025]. Available from: http://www.posit.co/
12. Christensen R. Ordinal-regression models for ordinal data. R Package Version 2023.12-4.1 [accessed February 7, 2025]. Available from: https://CRAN.r-project.org/package=ordinal
13. Lenth R. Emmeans: estimated marginal means, aka least-squares means. R Package Version 1.10.6 [accessed February 7, 2025]. Available from: https://CRAN.r-project.org/package=emmeans
14. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess 1994; 6: 284–290. https://doi.org/10.1037/1040-3590.6.4.284.
15. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016; 15: 155–163. https://doi.org/10.1016/j.jcm.2016.02.012.
16. Simpson E, Bissonnette R, Eichenfield LF, Guttman-Yassky E, King B, Silverberg JI, et al. The validated investigator global assessment for atopic dermatitis (VIGA-AD): the development and reliability testing of a novel clinical outcome measurement instrument for the severity of atopic dermatitis. J Am Acad Dermatol 2020; 83: 839–846. https://doi.org/10.1016/j.jaad.2020.04.104
17. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012; 22: 276–282.
18. Schmitt J, Spuls PI, Thomas KS, Simpson E, Furue M, Deckert S, 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: 800–807. https://doi.org/10.1016/j.jaci.2014.07.043
19. Oranje AP. Practical issues on interpretation of scoring atopic dermatitis: SCORAD index, objective SCORAD, patient-oriented SCORAD and three-item severity score. Curr Probl Dermatol 2011; 41: 149–155. https://doi.org/10.1159/000323308
20. Zhao CY, Tran AQT, Lazo-Dizon JP, Kim J, Daniel BS, Venugopal SS, et al. A pilot comparison study of four clinician-rated atopic dermatitis severity scales. Br J Dermatol 2015; 173: 488–497. https://doi.org/10.1111/bjd.13846
21. Carel K, Bratton DL, Miyazawa N, Gyorkos E, Kelsay K, Bender B, et al. The Atopic Dermatitis Quickscore (ADQ): validation of a new parent-administered atopic dermatitis scoring tool. Ann Allergy Asthma Immunol 2008; 101: 500–507. https://doi.org/10.1016/S1081-1206(10)60289-X
22. Charman CR, Venn AJ, Williams H, Bigby M. Measuring atopic eczema severity visually. Arch Dermatol 2005; 141. https://doi.org/10.1001/archderm.141.9.1146
23. Hanifin JM, Thurston M, Omoto M, Cherill R, Tofte SJ, Graeber M, et al. The Eczema Area and Severity Index (EASI): assessment of reliability in atopic dermatitis. Exp Dermatol 2001; 10: 11–18. https://doi.org/10.1034/j.1600-0625.2001.100102.x
24. Bożek A, Reich A. Assessment of intra- and inter-rater reliability of three methods for measuring atopic dermatitis severity: EASI, objective SCORAD, and IGA. Dermatology 2017; 233: 16–22. https://doi.org/10.1159/000472711
25. Davis CM, Flohr C, Gupta MR, Koplin JJ. Managing atopic dermatitis in patients with skin of color. J Allergy Clin Immunol Pract 2023; 11: 1376–1383. https://doi.org/10.1016/j.jaip.2023.03.041