

DR JONATHAN I SILVERBERG (Orcid ID : 0000-0003-3686-7805)

Article type : Original Article

Title: Validation of five patient-reported outcomes for atopic dermatitis severity in adults.

Running Header: PRO in adult AD.

J. I. Silverberg¹, D. J. Margolis², M. Boguniewicz³, L. Fonacier⁴, M. H. Grayson⁵, P. Y. Ong⁶,
Z. C. Fuxench², E. L. Simpson⁷

1. Northwestern University Feinberg School of Medicine, Chicago, IL
2. University of Pennsylvania Perelman School of Medicine, Philadelphia, PA
3. National Jewish Health and University of Colorado School of Medicine, Denver, CO
4. NYU Winthrop Hospital, Mineola, NY
5. Nationwide Children's Hospital–The Ohio State University College of Medicine, Columbus, OH
6. Children's Hospital Los Angeles, University of Southern California, Keck School of Medicine, Los Angeles, CA
7. Oregon Health & Science University, Portland, OR

Key Words: atopic dermatitis; eczema; severity; patient-reported outcomes; quality of life; burden; utility; mental health; physical health; signs; symptoms; validity

Abbreviations used: AD=atopic dermatitis; DLQI=Dermatology Life Quality Index; PRO = Patient-reported outcomes; QOL = quality of life; SF = Short Form; DLQI – Dermatology Life Quality Index; MCS = Mental Component Score; PCS = Physical Component Score; PO-SCORAD = Patient-Oriented Scoring Atopic Dermatitis; POEM = Patient-Oriented Eczema Measure

Corresponding author:

Jonathan I. Silverberg, MD, PhD, MPH
Northwestern University Feinberg School of Medicine
Department of Dermatology
Suite 1600

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.18002

This article is protected by copyright. All rights reserved.

676 N. St Clair St.
Chicago, IL 60611
Telephone: (312) 503-4958
Fax: (312) 695-0664
JonathanSilverberg@Gmail.com

Acknowledgements:

JI Silverberg had full access to all the data in the study and takes responsibility for the integrity of the data and accuracy of the data analysis.

Study concept and design: JI Silverberg

Acquisition of Data: Silverberg JI, Gelfand JM, Margolis D, Fonacier L, Boguniewicz M, Schwartz LB, Grayson MH, Ong P, Chiesa Fuxench Z

Analysis and interpretation of data: Silverberg JI, Gelfand JM, Margolis D, Fonacier L, Boguniewicz M, Schwartz LB, Grayson MH, Ong P, Chiesa Fuxench Z

Drafting of the manuscript: JI Silverberg

Critical revision of the manuscript for important intellectual content: Silverberg JI, Gelfand JM, Margolis D, Fonacier L, Boguniewicz M, Schwartz LB, Grayson MH, Ong P, Chiesa Fuxench Z

Statistical analysis: JI Silverberg

Administrative technical or material support: None

Study supervision: Allergy and Asthma Foundation of America

Financial disclosures:

Dr. Silverberg served as a consultant and/or advisory board member for Abbvie, Asana, Eli Lilly, Galderma, GlaxoSmithKline, Glenmark, Kiniksa, Leo, Menlo, Pfizer, Regeneron-Sanofi, Realm, Roivant receiving honoraria; speaker for Regeneron-Sanofi; and received research grants from GlaxoSmithKline and Regeneron-Sanofi. Dr. Silverberg is supported by the Dermatology Foundation.

Dr. Gelfand served as a consultant for BMS, Boehringer Ingelheim, GSK, Janssen Biologics, Menlo Therapeutics, Novartis Corp, Regeneron, Dr Reddy's labs, UCB (DSMB), Sanofi and Pfizer Inc., receiving honoraria; and receives research grants (to the Trustees of the University of Pennsylvania) from Abbvie, Janssen, Novartis Corp, Sanofi, Celgene, Ortho Dermatologics, and Pfizer Inc.; and received payment for continuing medical education work related to psoriasis that was supported indirectly by Lilly and Ortho Dermatologics. Dr. Gelfand is a co-patent holder of resiquimod for treatment of cutaneous T cell lymphoma. Dr. Gelfand is a Deputy Editor for the Journal of Investigative Dermatology receiving honoraria from the Society for Investigative Dermatology.

Dr. Chiesa Fuxench has served as a consultant for the National Eczema Association and the Asthma and Allergy Foundation, receiving honoraria, and receives or has received research grants (to the Trustees of the University of Pennsylvania) from Regeneron, Sanofi, Tioga and Vanda pharmaceuticals and Realm Therapeutics for work in atopic dermatitis; and has received payment for continuing medical education work related to atopic dermatitis that was supported indirectly by Regeneron and/or Sanofi.

Dr. Margolis is the chair of the data monitoring committee for all Sunovion clinical trials of Dupilumab, and has received independent research funding to his institution from the National Institute of Health and Valeant

Dr. Boguniewicz has received research funding from Anacor and Regeneron and consulted for Regeneron, Sanofi-Genzyme and Pfizer.

Dr. Fonacier has served as consultant for Regeneron, receiving honoraria; speaker for Regeneron; received research and educational grants from Genentech, Baxter, and Pfizer

Dr Grayson is a board member of AAFA and chair for the AAFA Medical Scientific Council, and has been on advisory boards for AstraZeneca, Genentech, and Novartis. Dr Simpson has served as a consultant and/or advisory board member for Regeneron-Sanofi. Dr. Ong is a co-investigator of Atopic Dermatitis Research Network and has consulted for Pfizer and Theravance.

Funding Support: Atopic Dermatitis in America Study is an independent research project of the Asthma and Allergy Foundation of America in partnership with the National Eczema Association (NEA) and sponsored by Sanofi Genzyme and Regeneron.

Funding/Sponsor was involved? No

Design and conduct of the study: Yes__ No_X_

Collection, management, analysis and interpretation of data: Yes__ No_X_

Preparation, review, or approval of the manuscript: Yes__ No_X_

Decision to submit the manuscript for publication: Yes__ No_X_

What's already known about this topic?

- Few studies demonstrated the validity of the atopic dermatitis severity assessments PO-SCORAD, PO-SCORAD sub-scores, NRS-itch and POEM.

What does this study add?

- This study demonstrated that PO-SCORAD, PO-SCORAD sub-scores, NRS-itch and POEM all had good construct validity in the assessment of atopic dermatitis severity in adults. Only POEM had observed floor effects.
- PO-SCORAD, PO-SCORAD sub-scores, NRS-itch and POEM all appear to have sufficient validity to be used as assessments of atopic dermatitis severity in clinical practice.

Abstract

Background: Structured patient-reported outcomes of AD severity are not standardized in clinical practice. We sought to determine the construct validity, internal consistency, cross-cultural validity, floor or ceiling effects of multiple AD severity assessments.

Methods: A cross-sectional, population-based study of 2893 adults, including 602 adults that met a modified UK Diagnostic Criteria for AD. AD severity was assessed using self-reported global AD severity, Patient-Oriented Eczema Measure (POEM), Patient-Oriented Scoring AD (PO-SCORAD) and its objective and subjective components, and numerical rating scale (NRS)-itch. QOL was assessed using short-form (SF-)12 mental and physical health scores, SF-6D health utility scores, Dermatology Life Quality Index (DLQI). Hospital Anxiety and Depression Scale (HADS) was assessed.

Results: PO-SCORAD, PO-SCORAD objective and subjective sub-scores, NRS-itch and POEM all had moderate to strong correlations with each other, and DLQI, fair to moderate correlations with HADS-A and HADS-D, and inverse correlations with SF-12 MCS and SF-6D (Pearson correlations, $P < 0.0001$). All scores showed good criterion validity as judged by analysis of variance and receiver operator characteristics. PO-SCORAD, PO-SCORAD objective sub-scores and POEM had similarly good internal consistency (Cronbach's $\alpha = 0.84, 0.82$ and 0.86); PO-SCORAD subjective sub-score was less internally consistent ($\alpha = 0.57$). All scores showed potentially poor cross-cultural validity as demonstrated by

uniform and non-uniform differential item functioning by age, sex and/or race/ethnicity for multiple items. There were floor effects for POEM, but not for the other assessments

Conclusions: PO-SCORAD, PO-SCORAD objective and subjective sub-scores, NRS-itch and POEM appear to be valid for assessing AD severity in clinical practice.

Introduction

Atopic Dermatitis (AD) is a heterogeneous disorder associated with a wide array of signs (e.g. lichenification, xerosis, exudative lesions, prurigo nodules)¹, distributions (e.g. flexural, extensor, hands, head and neck)^{1,2}, and symptoms (e.g. itch, pain³, sleep disturbance⁴⁻⁸, symptoms of depression, anxiety and inattentiveness^{5,7,9,10}), all of which contribute toward overall severity of disease^{11,12}. Numerous standardized AD severity assessments exist, with varying measurement constructs, validity and reliability.

Recently, the Harmonizing Outcome Measures in Eczema (HOME) international consensus group identified erythema, edema/papulation, excoriation and lichenification as four essential signs to be assessed in AD trials¹³. The HOME group recommended the Eczema Area and Severity Index (EASI) as the preferred outcomes assessments for AD signs in clinical trials¹⁴. Subsequently, Patient Oriented Eczema Measure (POEM) was selected as the preferred assessment for AD symptoms in clinical trials based on the available data to support their validity and reliability¹⁵. The HOME group is now deliberating which assessments should be recommended for use in clinical practice.

Patient-reported outcomes (PRO) are important to consider for measuring AD severity in clinical practice. POEM was derived from patients determining what symptoms caused them the most bother and assesses the frequency of 7 different patient-reported AD signs and symptoms¹⁶. The measurement properties of POEM have been examined in multiple studies and found good content and construct validity, but only fair internal consistency; cross-cultural validity has not been examined^{16,17}. Patient Oriented SCORAD (PO-SCORAD) is

another assessment of AD symptoms, which was designed as a PRO analogue of the SCORAD^{18,19}. Analogous to the clinician/investigator reported SCORAD, PO-SCORAD has patients assess the severity of 6 different signs of AD, lesional extent, with patient-reported intensity of itch and sleep disturbance being identical between both measures. PO-SCORAD has been found to correlate well with SCORAD¹⁸, and been shown to have some aspects of construct validity in a few studies²⁰⁻²²; discriminant validity, floor/ceiling effects, internal consistency and cross-cultural validity were not assessed. Some studies suggested that the objective (lesional severity and extent) or subjective (severity of itch and sleep disturbance) sub-scores of PO-SCORAD may also be valid. Previous studies found mixed results with respect to construct validity of subjective SCORAD/PO-SCORAD; discriminant validity, floor/ceiling effects, internal consistency and cross-cultural validity were not assessed^{14,16,23}.

More studies are needed to better demonstrate the validity and measurement properties of PO-SCORAD and its sub-scores. Further, few studies validated the numerical rating scale (NRS) itch in AD, despite its widespread use in clinical practice and trials. Finally, little is known about how the measurement properties of these different PRO compare with each other. In this study, we sought to determine the construct validity, internal consistency, cross-cultural validity, floor or ceiling effects of PO-SCORAD, PO-SCORAD sub-scores, NRS-itch and POEM, and compare their measurement properties with each other in US adults with AD.

Methods

Data Source

Data were obtained from the *AD in America* survey whose population was sampled from the long-standing GfK Knowledge Panel²⁴. The *AD in America* study was designed to examine multiple facets of the patient-burden, comorbidities and resource utilization of US

adults with AD. The GfK Knowledge panel is the largest and oldest probability-based web panel in the US and contains between 40,000 to 50,000 adult panel members at any given time. The GfK web panel was initially constructed from a national address-based sample of households in the US who are recruited to join in and receive small incentives for participating in web surveys on a regular basis. This approach uses a single sampling frame via the Delivery Sequence File of the US Postal Service to provide a statistically valid representation of the US population as well as many difficult-to-survey populations. The GfK web panel also provides internet access to households without existing internet access. This web-based panel has been previously used in other large, epidemiological studies and has been shown to be representative of the US population²⁵⁻²⁷. The survey questionnaire and protocol were approved by the ICF Institutional Review Board.

Study design

This was a cross-sectional study involving a two-stage sampling process. Stage 1 was designed to determine the prevalence of AD in US adults. In this stage, an initial cross-sectional sample of 2,137 adults from the existing GfK Knowledge web panel was invited to participate in the survey. The focus of the survey was not disclosed in the invitation to members of the web panel to avoid biasing participation based on respondent interest or disinterest in the subject. A total of 1,278 adults completed the survey (completion rate=60.2%). Although this sample provided a precise estimate of the prevalence of AD among the adult population, it did not yield a large enough sample of AD patients to investigate differences between different levels of disease severity. In stage 2, an additional sample of 8,217 adults from the GfK knowledge panel completed screening to identify and interview an additional group of adults with AD; 19.7% met UKWP criteria and were eligible for inclusion (completion rate=59.9%) (Supplemental Figure 1). Using data from the US

Census Bureau, sample weights were created that adjusted for age, gender, race, ethnicity, education level, census region, household income, home ownership status and metropolitan area using an iterative proportional fitting procedure. Sample weights were included in all analyses to allow for representative estimates of the US population.

AD and AD severity assessments

An adaptation of the United Kingdom Working Party (UKWP) criteria was selected by the *AD in America* advisory committee as the screening tool for patient eligibility²⁸. This included all aspects of the UKWP criteria (having an itchy skin condition during the past 12 months and 3 or more of the following: [1] history of skin crease involvement; [2] a personal history of asthma or hay fever; [3] a history of general dry skin during the past year; [4] onset under the age of two), except that assessment of visible flexural eczema by a clinician was not performed.

Self-assessments of AD severity were completed in the following order for logistic reasons: 1. short-form 12 (SF-12) (12 items, 8 domains, using mental component score [MCS], physical component scores [PCS], and Brazier's SF-6D scoring)²⁹⁻³²; 2. the previously validated self-reported global AD severity question "Would you describe your atopic dermatitis or eczema as mild, moderate, or severe?"¹¹; 3. Patient-Oriented Scoring AD (PO-SCORAD) index [9 items; range: 0-103], PO-SCORAD objective [7 items; range: 0-83] and subjective [2 items; range: 0-20]³³ sub-scores; 4. NRS-itch; 5. Patient-Oriented Eczema Measure (POEM) [7 items; range: 0-28]³⁴; and 6. Hospital Anxiety and Depression Scale, anxiety (HADS-A) and depression (HADS-D) scores (7 items each; range: 0-21)^{35,36}. These patient-reported outcomes (PRO) were selected *a priori* based on their feasibility, being

preferred measures by international consensus of the HOME group^{15,37} and/or previously being validated^{11,12,15,37}.

Statistical analysis

Summary statistics were estimated for baseline population characteristics. Convergent construct validity of PO-SCORAD, PO-SCORAD objective and subjective sub-scores, NRS-itch, and POEM were established using Pearson correlations with DLQI, SF-12 MCS, PCS and SF-6D, HADS-A and HADS-D scores. Since POEM is the preferred assessment for AD symptoms in clinical trials by the HOME group, we compared the strength of correlations of PO-SCORAD, PO-SCORAD objective and subjective sub-scores and NRS-itch with POEM as the reference assessment using Fisher's test of z-scores. Criterion validity was determined using analysis of variance. In addition, logistic regression models were constructed with self-reported global AD severity as the binary dependent variable (moderate vs. mild, severe vs. moderate). The independent variables were PO-SCORAD, PO-SCORAD objective and subjective sub-scores, NRS-itch and POEM scores. Discriminant validity was determined using area under the curve (AUC). AUC scores of ≥ 0.90 were considered excellent, 0.80–0.89 good, 0.70–0.79 fair, <0.70 poor and <0.60 fail³⁸. Ordinal logistic regression was not used because the data did not meet the proportional odds assumption (Score test, $P < 0.01$).

Internal consistency was determined using Cronbach's alpha and Pearson correlation between individual items. Floor or ceiling effects of total scores and individual items were considered present if 15% of responses fell in the lowest or highest scores^{39,40}. The above statistical analyses were performed in SAS version 9.4.3 and sample weights were incorporated into all analyses (SAS Institute, Cary, IN). Missing values were encountered in $\leq 0.1\%$ of respondents for all analyzed variables. Complete case analysis was performed, i.e.

respondents with missing values were excluded from analysis. A two-sided P-value of 0.05 was considered statistically significant.

Cross-cultural item validity was assessed using differential item functioning (DIF) across respondents with AD by sex (male/female), race (white/black/Asian/other) and age (<50/≥50 years). DIF occurs when respondents from different groups have differing probabilities of success on an item, after controlling for overall ability based on their total scores. If DIF is present, then item responses may be related to both the underlying trait being measured as well as other factors. DIF was detected for polytomous items, using an iterative hybrid of ordinal logistic regression and item response theory (lordif-package, R version 3.5.1, Vienna, Austria)⁴¹.

Results

Patient characteristics

Overall, 2,893 adults were included in this study; 602 met AD criteria (prevalence [95% CI]: 7.39% [5.81-8.97%]). Subjects with AD were 53.6% female and 71.9% Caucasian/white, with a mean ± std. dev. age of 52.0±16.3 years. Among persons with AD, the weighted mean [95% CI] PO-SCORAD was 27.5 [25.7-29.3], PO-SCORAD objective and subjective sub-scores were 17.5 [16.4-18.5] and 8.4 [8.0-8.8], NRS-itch was 4.5 [4.2-4.8], POEM was 7.5 [6.8-8.1], and DLQI was 4.9 [4.2-5.5].

Convergent validity

Regarding convergent validity, PO-SCORAD, PO-SCORAD objective and subjective sub-scores, NRS-itch and POEM had moderate to strong correlations with each other, and DLQI, and fair to moderate correlations with HADS-A and HADS-D and inverse correlations

with SF-12 MCS and SF-6D, but only weak inverse correlations with SF-12 PCS (Pearson correlations, $P < 0.0001$) (Table 1).

Compared to POEM, PO-SCORAD had stronger correlations with DLQI ($P = 0.003$), SF-12 MCS ($P = 0.002$), SF-6D ($P = 0.006$), HADS-A ($P = 0.0001$) and HADS-D ($P = 0.005$), but not SF-12 PCS ($P = 0.57$). Whereas, PO-SCORAD objective sub-score had stronger correlations than POEM with DLQI ($P = 0.02$), but no significant differences for MCS ($P = 0.10$), PCS ($P = 0.19$), or SF-6D ($P = 0.08$). PO-SCORAD subjective sub-score had stronger correlations with DLQI ($P = 0.04$), MCS ($P < 0.0001$), SF-6D ($P = 0.005$), but not PCS ($P = 0.39$). NRS-itch had significantly weaker correlations than POEM with DLQI ($P = 0.0006$), but not with MCS ($P = 0.63$), PCS ($P = 0.87$), SF-6D ($P = 0.43$), HADS-A ($P = 0.93$) or HADS-D ($P = 0.94$).

Criterion validity

There is no gold-standard assessment for AD severity. Thus, we compared PO-SCORAD and POEM with self-reported global AD severity to examine criterion validity. There were significant and stepwise (monotonic) increases of PO-SCORAD, PO-SCORAD objective and subjective sub-scores, NRS-itch, and POEM with each level of self-reported global AD severity (ANOVA, $P < 0.0001$ for all) (Figure 1).

PO-SCORAD, PO-SCORAD objective and subjective sub-scores, NRS-itch and POEM scores also showed good criterion validity in predicting overall AD severity (Table 2). The AUC for the five measures were good to excellent for distinguishing severe vs. mild disease, and good for distinguishing moderate vs. mild or severe vs. moderate AD, indicating that they all had good ability to distinguish between severity groups.

Internal consistency

PO-SCORAD, PO-SCORAD objective and subjective sub-score, and POEM items showed good internal consistency among adults with AD (Cronbach's alpha: 0.84, 0.82, 0.57, and 0.86, respectively), with weak to moderate inter-item correlations (Supplemental Table 1).

Cross-cultural validity

Significant uniform and non-uniform DIF were found for multiple PO-SCORAD and POEM items and/or NRS-itch by sex, race/ethnicity and age (Supplemental Table 2).

Floor or ceiling effects

The proportions of patients with lowest and highest values for PO-SCORAD (1.5% or 0.2%), PO-SCORAD objective (5.9% or 0.2%) and subjective sub-scores (2.4% and 0.8%) and NRS-itch (6.7% and 2.7%) were below 15%, indicating there were no floor or ceiling effects overall. However, there were floor (15.1%) but no ceiling (0.7%) effects observed for POEM.

Multiple items for PO-SCORAD (redness, swelling, scabs/oozing, scratching, thickening and NRS-sleep) and all items for POEM individually had floor effects; no individual items had ceiling effects.

Discussion

This study demonstrated that PO-SCORAD, the objective and subjective sub-scores of PO-SCORAD, NRS-itch and POEM all had good convergent validity for QOL impact and mental health symptoms, criterion validity for AD severity, and good internal consistency. PO-SCORAD, PO-SCORAD objective and subjective sub-scores showed significantly better correlations with QOL and mental health assessments compared to POEM. Whereas, NRS-itch compared to POEM had significantly lower correlations with DLQI, but similar results with other QOL and mental health scores. Of note, all of the AD severity scores studied were only weakly or moderately correlated with QOL and mental health symptoms, suggesting that they each measure different disease-related domains. PO-SCORAD, PO-SCORAD objective and subjective sub-scores and NRS-itch did not have any significant floor or ceiling effects, whereas POEM had significant floor effects. Floor effects raise potential concern about how well POEM may perform in clinical trials and/or practice of patients with milder AD. The apparent floor effects may be due to a large proportion of US adults with AD only having very mild disease. However, this scenario is unlikely given the absence of floor-effects observed for any other assessment. Moreover, floor effects were observed in 5 of the 7 items in POEM in the original validation study conducted in a hospital dermatology clinic¹⁶. Together, the results show that PO-SCORAD, PO-SCORAD objective and subjective sub-scores, NRS-itch and POEM are sufficiently valid for the assessment of AD severity in adults, and may be a valuable assessment in clinical practice and epidemiological studies.

While PO-SCORAD, PO-SCORAD sub-scores, NRS-itch and POEM all are valid assessments of AD severity, they each measure different constructs. POEM assesses the frequency of AD symptoms, which generally correlates with severity of AD signs and symptoms. However, the present study suggests that PO-SCORAD, PO-SCORAD objective and subjective sub-scores that directly assess the severity of AD signs and symptoms may

Accepted Article

better predict the burden of AD, since PO-SCORAD had significantly stronger correlations with multiple QOL and HADS domains and numerically greater AUC than POEM when compared to global AD severity. This is an important consideration for use in clinical practice. All five assessments are feasible, but previous studies found shorter completion times for POEM (1-2 minutes), NRS-itch (not formally assessed, but only 1 question and expected to be <1 minutes) and subjective PO-SCORAD (not formally assessed, but only 2 questions and expected to be 1-2 minutes) than PO-SCORAD (<5 or <10 minutes)²³. NRS-itch is scored as a single item ranging from 0-10. PO-SCORAD subjective sub-scores is a sum of 2 items, each on a range of 0-10. POEM is also simple to score as the sum of 7 items, each on a range of 0 to 4. Whereas, PO-SCORAD requires summing of items on different scales and weighing scores from different components to obtain a total score. It would be more difficult to hand-score PO-SCORAD or PO-SCORAD objective sub-score in clinical practice, though apps and other tools may simplify this task. Another simple and feasible assessment that can potentially be used in clinical practice is self-reported global AD severity, which was previously found to have good content and construct validity^{11,12} and encompasses is best predicted by a combination of PO-SCORAD, POEM, NRS-itch, NRS-pain, and HADS ¹².

Both PO-SCORAD, NRS-itch and POEM had multiple items that suffered from uniform and/or non-uniform DIF by age, sex and race/ethnicity, indicating potentially poor cross-cultural validity. Items with DIF show different correlations with the overall scores between different groups. That is, the items with DIF performed differently across various demographic subgroups, which is not merely due to between group differences. Thus, scores for these different patient groups should neither be combined nor compared during analysis of clinical trials or research studies. While this may not be practical in clinical trials, at the very least, scores for these measures should also be presented stratified by the various

demographic groups. In addition, other instruments might be considered, which do not suffer from DIF.

These results are consistent with previous findings of moderate to strong correlations of PO-SCORAD and the objective sub-score of PO-SCORAD with POEM and QOL in patients^{20,21} and families²². PO-SCORAD and PO-SCORAD objective sub-score were also found to have moderate to strong correlations with SCORAD^{18,21} and objective SCORAD²⁰, respectively. Similarly, previous studies found good correlations of POEM with objective AD severity and QOL assessments⁴².

This study has several strengths, including use of a large, nationally representative cohort, and good representation across gender, race/ethnicity and AD severity. However, there are some limitations. We did not assess responsiveness, test-retest reliability or content validity, which are all important considerations in the choice of instruments. However, many of these properties were assessed in previous studies. POEM demonstrated good content validity, fair responsiveness, but poor reliability^{14,16,23}. PO-SCORAD demonstrated fair responsiveness, but reliability has not been examined²³. Subjective PO-SCORAD demonstrated poor content validity and responsiveness, but reliability was not examined²³. All severity assessments were based on PRO, and we were unable to compare PO-SCORAD against the SCORAD. Multiple PRO were completed in a standardized order. It is possible that the initial completion of one PRO may have subconsciously influenced responses to the following PRO. Self-reported global severity was used as the gold-standard comparator for establishing criterion validity. Only three response options were used (mild, moderate or severe) based on previous studies demonstrating content and construct validity of this question¹². Selection and/or non-response bias may be present. An adaptation of the UKWP criteria was used that did not include physical examination for flexural eczema. This approach might have slightly over-estimated AD prevalence. This study used a

epidemiological cohort, and the results may have differed in a clinical dermatology setting.

Finally, we did not assess the measurement properties of these PRO in children. Additional studies are needed to confirm the validity of these PRO in children and across different settings.

In conclusion, the present study demonstrates that PO-SCORAD, PO-SCORAD objective and subjective sub-scores, NRS-itch and POEM all have good construct validity and may be useful for the assessment of AD severity in clinical practice and epidemiological research. Additional research is needed to directly confirm these results in the clinical practice setting and to assess content validity and reliability.

Figure Legends

Supplemental Figure 1. Design of the AD in America study.

Figure 1. Criterion validity of PO-SCORAD, PO-SCORAD objective and subjective sub-scores, NRS-itch and POEM. Box-whisker plots and overlaid jitter plots of PO-SCORAD, PO-SCORAD objective and subjective sub-scores, NRS-itch and POEM scores are presented for each severity level of self-reported global AD severity.

References

- 1 Yew YW, Thyssen JP, Silverberg JI. A systematic review and meta-analysis of the regional and age-related differences in atopic dermatitis clinical characteristics. *Journal of the American Academy of Dermatology* 2019; **80**: 390-401.
- 2 Silverberg JI, Margolis DJ, Boguniewicz M *et al.* Distribution of atopic dermatitis lesions in United States adults. *Journal of the European Academy of Dermatology and Venereology : JEADV* 2019.
- 3 Vakharia PP, Chopra R, Sacotte R *et al.* Burden of skin pain in atopic dermatitis. *Annals of allergy, asthma & immunology : official publication of the American College of Allergy, Asthma, & Immunology* 2017; **119**: 548-52 e3.
- 4 Fishbein AB, Vitaterna O, Haugh IM *et al.* Nocturnal eczema: Review of sleep and circadian rhythms in children with atopic dermatitis and future research directions. *The Journal of allergy and clinical immunology* 2015; **136**: 1170-7.
- 5 Silverberg JI. Selected comorbidities of atopic dermatitis: Atopy, neuropsychiatric, and musculoskeletal disorders. *Clin Dermatol* 2017; **35**: 360-6.
- 6 Silverberg JI, Garg NK, Paller AS *et al.* Sleep disturbances in adults with eczema are associated with impaired overall health: a US population-based study. *The Journal of investigative dermatology* 2015; **135**: 56-66.
- 7 Silverberg JI, Simpson EL. Association between severe eczema in children and multiple comorbid conditions and increased healthcare utilization. *Pediatric allergy and immunology : official publication of the European Society of Pediatric Allergy and Immunology* 2013; **24**: 476-86.
- 8 Yu SH, Attarian H, Zee P *et al.* Burden of Sleep and Fatigue in US Adults With Atopic Dermatitis. *Dermatitis : contact, atopic, occupational, drug : official journal of the American Contact Dermatitis Society, North American Contact Dermatitis Group* 2016; **27**: 50-8.
- 9 Patel KR, Immaneni S, Singam V *et al.* Association between atopic dermatitis, depression and suicidal ideation: A systematic review and meta-analysis. *Journal of the American Academy of Dermatology* 2018.
- 10 Yu SH, Silverberg JI. Association between Atopic Dermatitis and Depression in US Adults. *The Journal of investigative dermatology* 2015; **135**: 3183-6.
- 11 Vakharia PP, Chopra R, Sacotte R *et al.* Validation of patient-reported global severity of atopic dermatitis in adults. *Allergy* 2018; **73**: 451-8.
- 12 Silverberg JI, Chiesa Fuxench ZC, Gelfand JM *et al.* Content and construct validity, predictors, and distribution of self-reported atopic dermatitis severity in US adults. *Annals of allergy, asthma & immunology : official publication of the American College of Allergy, Asthma, & Immunology* 2018; **121**: 729-34 e4.
- 13 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; **171**: 1318-25.
- 14 Schmitt J, Langan S, Williams HC *et al.* What are the best outcome measurements for atopic eczema? A systematic review. *The Journal of allergy and clinical immunology* 2007; **120**: 1389-98.
- 15 Spuls PI, Gerbens LA, Simpson E *et al.* POEM a core instrument to measure symptoms in clinical trials: a HOME statement. *The British journal of dermatology* 2016.
- 16 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. *Archives of dermatology* 2004; **140**: 1513-9.

- 17 Schmitt J, Langan S, Williams HC. What are the best outcome measurements for
atopic eczema? A systematic review. *J Allergy Clin Immunol* 2007; **120**: 1389-98.
- 18 Stalder JF, Barbarot S, Wollenberg A *et al*. Patient-Oriented SCORAD (PO-
SCORAD): a new self-assessment scale in atopic dermatitis validated in Europe.
Allergy 2011; **66**: 1114-21.
- 19 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*
2009; **218**: 246-51.
- 20 van Oosterhout M, Janmohamed SR, Spierings M *et al*. Correlation between
Objective SCORAD and Three-Item Severity Score used by physicians and Objective
PO-SCORAD used by parents/patients in children with atopic dermatitis.
Dermatology 2015; **230**: 105-12.
- 21 Boccardi D, D'Auria E, Turati F *et al*. Disease severity and quality of life in children
with atopic dermatitis: PO-SCORAD in clinical practice. *Minerva Pediatr* 2017; **69**:
373-80.
- 22 Pustisek N, Vurnek Zivkovic M, Situm M. Quality of Life in Families with Children
with Atopic Dermatitis. *Pediatric dermatology* 2016; **33**: 28-32.
- 23 Gerbens LA, Prinsen CA, Chalmers JR *et al*. Evaluation of the measurement
properties of symptom measurement instruments for atopic eczema: a systematic
review. *Allergy* 2017; **72**: 146-63.
- 24 Chiesa Fuxench ZC, Block JK, Boguniewicz M *et al*. Atopic Dermatitis in America
Study: A Cross-Sectional Study Examining the Prevalence and Disease Burden of
Atopic Dermatitis in the US Adult Population. *The Journal of investigative*
dermatology 2019; **139**: 583-90.
- 25 Rowen TS, Gaither TW, Awad MA *et al*. Pubic Hair Grooming Prevalence and
Motivation Among Women in the United States. *JAMA dermatology* 2016; **152**:
1106-13.
- 26 Hays RD, Liu H, Kapteyn A. Use of Internet panels to conduct surveys. *Behavior*
research methods 2015; **47**: 685-90.
- 27 Heckman CJ, Darlow S, Manne SL *et al*. Correspondence and correlates of couples'
skin cancer screening. *JAMA dermatology* 2013; **149**: 825-30.
- 28 Williams HC, Burney PG, Pembroke AC *et al*. The U.K. Working Party's Diagnostic
Criteria for Atopic Dermatitis. III. Independent hospital validation. *The British*
journal of dermatology 1994; **131**: 406-16.
- 29 Jenkinson C, Layte R. Development and testing of the UK SF-12 (short form health
survey). *Journal of health services research & policy* 1997; **2**: 14-8.
- 30 Ware J, Jr., Kosinski M, Keller SD. A 12-Item Short-Form Health Survey:
construction of scales and preliminary tests of reliability and validity. *Medical care*
1996; **34**: 220-33.
- 31 Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of
health from the SF-36. *Journal of health economics* 2002; **21**: 271-92.
- 32 Silverberg JI, Gelfand JM, Margolis DJ *et al*. Validation and interpretation of short
form 12 and comparison with dermatology life quality index in atopic dermatitis in
adults. *J Invest Derm* 2019; **In Press**.
- 33 Stalder JF, Barbarot S, Wollenberg A *et al*. Patient-Oriented SCORAD (PO-
SCORAD): a new self-assessment scale in atopic dermatitis validated in Europe.
Allergy 2011; **66**: 1114-21.
- 34 Charman CR, Venn AJ, Ravenscroft JC *et al*. 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.

- 35 Wittkowski A, Richards HL, Griffiths CE *et al*. The impact of psychological and
clinical factors on quality of life in individuals with atopic dermatitis. *J Psychosom*
36 *Res* 2004; **57**: 195-200.
- 37 Zigmond AS, Snaith RP. The hospital anxiety and depression scale. *Acta Psychiatr*
Scand 1983; **67**: 361-70.
- 38 Chalmers JR, Thomas KS, Apfelbacher C *et al*. Report from the fifth international
consensus meeting to harmonize core outcome measures for atopic eczema/dermatitis
clinical trials (HOME initiative). *The British journal of dermatology* 2018; **178**: e332-
e41.
- 39 Metz CE. Basic principles of ROC analysis. *Semin Nucl Med* 1978; **8**: 283-98.
- 40 Terwee CB, Bot SD, de Boer MR *et al*. Quality criteria were proposed for
measurement properties of health status questionnaires. *Journal of clinical*
epidemiology 2007; **60**: 34-42.
- 41 Lim CR, Harris K, Dawson J *et al*. Floor and ceiling effects in the OHS: an analysis
of the NHS PROMs data set. *BMJ open* 2015; **5**: e007765.
- 42 Choi SW, Gibbons LE, Crane PK. lordif: An R Package for Detecting Differential
Item Functioning Using Iterative Hybrid Ordinal Logistic Regression/Item Response
Theory and Monte Carlo Simulations. *J Stat Softw* 2011; **39**: 1-30.
- 43 Vakharia PP, Cella D, Silverberg JI. Patient-reported outcomes and quality of life
measures in atopic dermatitis. *Clin Dermatol* 2018; **36**: 616-30.

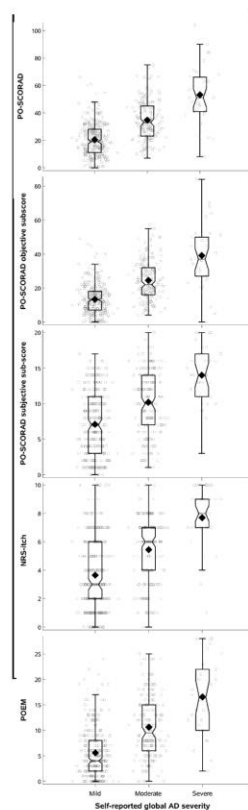
Table 1. Pearson correlation of PO-SCORAD, PO-SCORAD objective and subjective sub-scores, NRS-itch and POEM with each other, and measures of quality of life, anxiety and depression.										
Assessment	Pearson r									
	Atopic dermatitis symptoms				Quality of life				Mental health	
	PO-SCORAD objective sub-score	PO-SCORAD subjective sub-score	NRS- itch	POEM	DLQI	SF-12 MCS	SF-12 PCS	SF-6D	HADS-A	HADS-D
Any site (n=602)										
PO-SCORAD	0.97***	0.76***	0.71***	0.71***	0.71***	-0.47***	-0.12**	-0.44***	0.52***	0.44***
PO-SCORAD objective sub-score	1.00	0.59***	0.59***	0.69***	0.69***	-0.40***	-0.07	-0.39***	0.46***	0.39***
PO-SCORAD subjective sub-score		1.00	0.82***	0.55***	0.53***	-0.51***	-0.21***	-0.44***	0.51***	0.46***
NRS-itch			1.00	0.58***	0.48***	-0.34***	-0.14**	-0.31***	0.34***	0.31***
POEM				1.00	0.62***	-0.32***	-0.15***	-0.30***	0.33***	0.31***

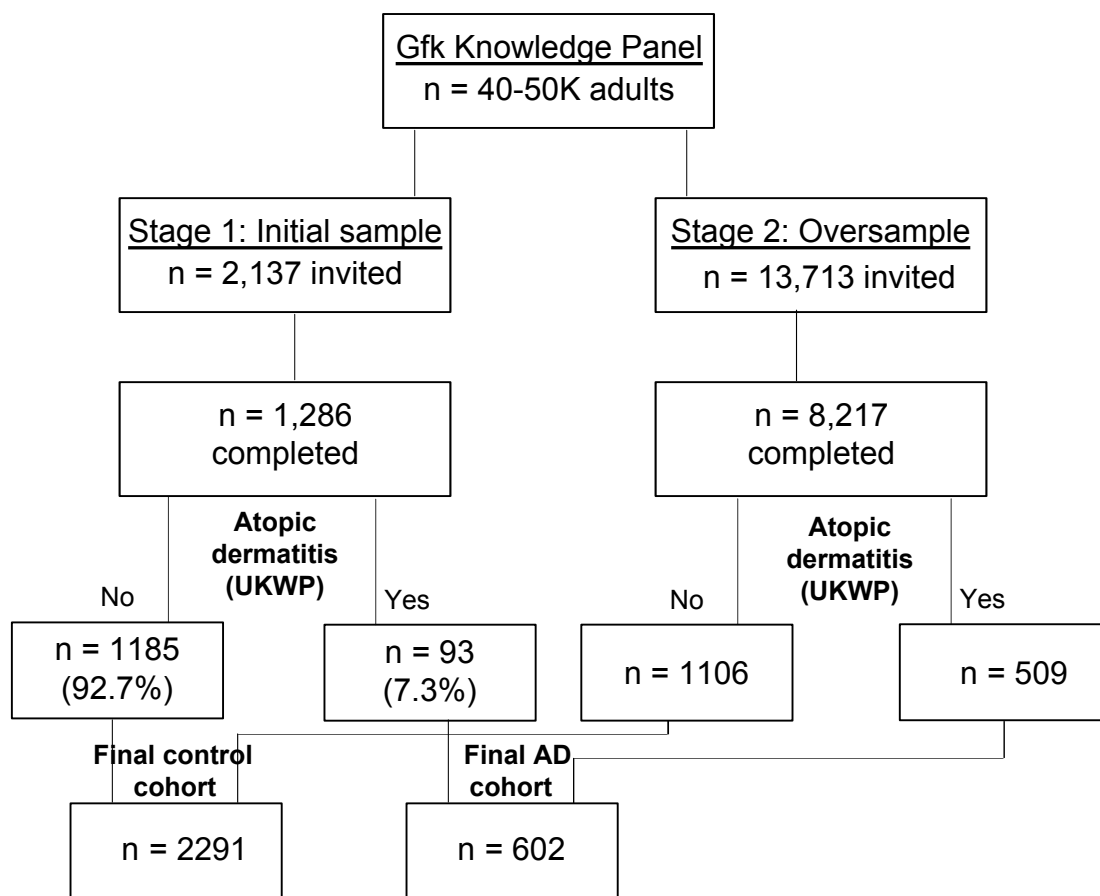
* P<0.05; ** P<0.01; *** P<0.0001

Table 2. Area under the ROC curve for PO-SCORAD, PO-SCORAD objective and subjective sub-scores, NRS-itch and POEM as predictors of overall AD severity.

Groups	AUC			NRS-itch	POEM
	PO-SCORAD	PO-SCORAD objective sub-score	PO-SCORAD Subjective sub-score		
Global severity					
Moderate vs. mild	0.78	0.78	0.69	0.71	0.75
Severe vs. mild	0.91	0.89	0.87	0.88	0.89
Severe vs. moderate	0.77	0.75	0.73	0.78	0.73

P<0.0001 for all





Supplemental Table 1. Inter-item correlations for PO-SCORAD, NRS-itch and POEM in US adults with atopic dermatitis.

PO-SCORAD item	Pearson r								
	Number of sites	Dry	Red areas	Swollen	Scabs or oozing	Scratch marks	Thickening	NRS itch	NRS sleep
Number of sites	1.00	0.19	0.23	0.28	0.21	0.35	0.31	0.44	0.30
Dry		1.00	0.35	0.40	0.29	0.32	0.41	0.38	0.22
Red areas			1.00	0.59	0.60	0.42	0.45	0.36	0.32
Swollen				1.00	0.65	0.55	0.53	0.38	0.37
Scabs or oozing					1.00	0.55	0.53	0.37	0.29
Scratch marks						1.00	0.56	0.50	0.30
Thickening							1.00	0.43	0.25
NRS itch								1.00	0.43
NRS sleep									1.00
POEM item	Itchy	Sleep	Bleeding	Weeping or oozing	Cracking	Flaking	Dry or rough		
Itchy	1.00	0.53	0.39	0.29	0.50	0.52	0.64		
Sleep		1.00	0.55	0.47	0.50	0.40	0.38		
Bleeding			1.00	0.72	0.51	0.45	0.33		
Weeping or cracking				1.00	0.45	0.37	0.25		
Cracking					1.00	0.70	0.63		
Flaking						1.00	0.66		
Dry or rough							1.00		

P<0.0001 for all correlations

Supplemental Table 2. Differential Item Functioning (DIF) by sex, race/ethnicity and age of items in PO-SCORAD, NRS-itch and POEM.

	Item	Adults with AD (n=602)					
		DIF by sex		DIF by race/ethnicity		DIF by age	
		Uniform	Non-uniform	Uniform	Non-uniform	Uniform	Non-uniform
PO-SCORAD	Number of sites	0.003	0.01	0.23	0.06	0.0001	0.0007
	Dry	0.004	0.05	0.0001	0.01	0.0002	0.001
	Red areas	0.64	0.70	0.08	0.0001	<0.0001	0.03
	Swollen	0.15	0.23	0.60	0.10	0.003	0.67
	Scabs or oozing	0.71	0.15	0.58	0.09	0.07	0.18
	Scratch marks	0.86	0.70	0.07	0.80	0.006	0.58
	Thickening	0.47	0.46	0.48	0.41	0.01	0.04
	NRS-itch	0.001	0.001	<0.0001	0.01	0.10	0.46
	NRS-sleep	0.003	0.92	0.0003	0.61	0.98	0.72
POEM	Itchy	0.04	0.03	0.45	0.95	0.0001	0.003
	Sleep	0.56	0.77	<0.0001	0.02	0.42	0.27
	Bleeding	0.61	0.27	<0.0001	0.62	<0.0001	0.41
	Weeping or	0.06	0.01	<0.0001	0.04	<0.0001	0.03
	Cracking	0.08	<0.0001	<0.0001	0.40	<0.0001	0.04
	Flaking	0.19	0.45	0.08	0.06	0.30	0.25
	Dry or rough	<0.0001	0.87	<0.0001	<0.0001	0.12	<0.0001

Supplemental Table 3. Distribution of responses for PO-SCORAD, NRS-itch, and POEM items.						
Items	Freq (%)					
PO-SCORAD						
	0	1-5	6-15	16-25	26-35	36
Number of sites	73 (12.1%)	298 (49.5%)	194 (32.2%)	27 (4.5%)	8 (1.3%)	2 (0.3%)
	Not at all	Slightly dry	Moderately dry	Extremely dry		
Dry	88 (14.7%)	221 (36.9%)	231 (38.6%)	59 (9.9%)		
	Not at all	Slightly red	Moderately red	Extremely red		
Red areas	313 (52.1%)	210 (34.9%)	62 (10.3%)	16 (2.7%)		
	Not at all	Slightly swollen	Moderately swollen	Extremely		
Swollen	447 (74.3%)	109 (18.1%)	36 (6.0%)	10 (1.7%)		
	Not at all	A few	A moderate number	Several		
Scabs or oozing	480 (79.7%)	100 (16.6%)	16 (2.7%)	6 (1.0%)		
	Not at all	A few scratch marks	A moderate number of scratch marks	Several scratch marks		
Scratch marks	278 (46.2%)	251 (41.7%)	56 (9.3%)	17 (2.8%)		
	Not at all thickening	Slight thickening	Moderate thickening	Significant thickening		
Thickening	339 (56.3%)	171 (28.4%)	71 (11.8%)	21 (3.5%)		
	0	1-3	4-6	7-9	10	
NRS-itch	40 (6.7%)	218 (36.5%)	186 (31.1%)	138 (23.1%)	16 (2.7%)	
NRS-sleep	121 (20.3%)	161 (27.0%)	138 (23.2%)	160 (26.9%)	16 (2.7%)	
POEM	No days	1-2 days	3-4 days	5-6 days	Every day	
Itchy	137 (22.8%)	186 (30.9%)	119 (19.8%)	37 (6.2%)	123 (20.4%)	
Sleep	357 (59.3%)	121 (20.1%)	67 (11.1%)	21 (3.5%)	36 (6.0%)	
Bleeding	449 (75.6%)	101 (16.8%)	28 (4.7%)	12 (2.0%)	12 (2.0%)	
Weeping or cracking	497 (82.6%)	51 (8.5%)	32 (5.3%)	9 (1.5%)	13 (2.2%)	
Cracking	334 (55.5%)	127 (21.1%)	52 (8.6%)	18 (3.0%)	70 (11.6%)	
Flaking	272 (45.2%)	172 (28.6%)	48 (8.0%)	25 (4.2%)	85 (14.1%)	
Dry or rough	139 (23.1%)	170 (28.2%)	87 (14.5%)	36 (6.0%)	170 (28.2%)	

Missing values were present in ≤6 subjects for any PO-SCORAD, 0 subjects for POEM questions, and 0 for sex, race/ethnicity and age. Complete case analysis was performed.