What are the best outcome measurements for atopic eczema? A systematic review

Jochen Schmitt, MD, MPH, Sinead Langan, MD, and Hywel C. Williams, PhD, FRCP, on behalf of the European Dermato-Epidemiology Network Dresden, Germany, and Nottingham, United Kingdom

Background: Valid and reliable outcome measurements are a prerequisite for evidence-based practice. The comparative validity and reliability of outcome measurements for assessing atopic eczema (AE) severity is unclear.

Objective: We sought to assess the validity, reliability, sensitivity to change, and ease of use of outcome measurements for AE. We also sought to give recommendations on which outcomes to use in clinical research and for clinical monitoring.

Methods: We performed a systematic review and survey of clinical experts and patients.

Results: Twenty published outcome measurements were identified. There is evidence of adequate construct validity for 3 measurements (Severity Scoring of Atopic Dermatitis index [SCORAD], Eczema Area and Severity Index [EASI], and Three Item Severity Score), adequate internal consistency of 1 scale (Patient-oriented Eczema Measure [POEM]), adequate interobserver reliability of 5 measurements (Basic Clinical Scoring System; Nottingham Eczema Severity Score; Objective Severity Assessment of Atopic Dermatitis; Six Area, Six Sign Atopic Dermatitis severity score; and SCORAD), adequate test-retest reliability of 1 scale (POEM), and adequate sensitivity to change of 3 measurements (EASI, SCORAD, and Investigators’ Global Atopic Dermatitis Assessment). Most outcome measurements have adequate content validity, as assessed by patients and experts. Data on the time to perform the assessment was identified for 8 outcome measurements. Only SCORAD, EASI, and POEM have been tested sufficiently and performed adequately.

Conclusion: There are too many published outcome measures for AE. Most have not been tested properly or perform adequately when tested, and their continued use hampers scientific communication.

Clinical implications: Only SCORAD, EASI, and POEM currently perform adequately. These scales should be used in future studies. (J Allergy Clin Immunol 2007;120: 1389-98.)

Key words: Atopic eczema, clinical research, evidence-based medicine, outcome measurement, reliability, severity of illness index, validity

Atopic eczema (AE; synonymous with atopic dermatitis) has a high prevalence, causes considerable morbidity, and imposes a high economic burden. Adequate laboratory tests to assess disease severity in randomized controlled trials (RCTs) and to monitor treatment in clinical practice do not exist. Therefore valid and reliable clinical outcome measures are a necessary prerequisite for good evidence-based practice. Summarizing our collective knowledge on therapeutic interventions through systematic reviews and meta-analyses can only be conducted when standardized and valid outcome measurements are consistently applied in RCTs.

In a systematic review of all RCTs of therapeutic interventions for AE published between 1994 and 2001, Charman et al found that only 27% of the investigators who incorporated an “objective” assessment of clinical severity as an outcome applied a severity scale that had been published before. The authors identified a total of 56 different objective measurements of disease severity in 94 trials (0.6 outcomes per trial). Another systematic review revealed that 7 different measurements of objective disease severity had been used in 14 trials on cyclosporine in AE.

Not only is there poor standardization of outcomes in RCTs, but there is also a great lack of validation of the measurement methods used. In 2000, Charman and Williams identified a total of 13 named outcome measurements of disease severity of AE and reported that the vast majority of the measurements identified had not been validated appropriately. The review by Charman and Williams focused on the question of whether published outcome measurements had been tested at all and not whether they performed sufficiently well when tested.

Given the importance of AE as a health issue both for individual patients and for health care systems, it is critically important to examine the validity and reliability of existing outcome measurements and then to find a consensus among those short-listed scales between all stakeholders, including patients, clinicians, researchers, drug companies, and regulatory authorities.

The objectives of this systematic review are to (1) update the abovementioned review on named measurements of objective disease severity of AE and (2) extend...
Assessment of psychometric properties from the literature

Before adopting an outcome measurement into clinical practice, it should be tested for reliability, validity, sensitivity to change, and acceptability.8 Validity concerns the soundness of the inferences based on a measurement (ie, whether it truly measures what it is supposed to), whereas reliability concerns the degree to which a measurement is assessed without random error.9 From the literature, we assessed construct validity, internal consistency, interobserver reliability, test-retest reliability, sensitivity to change, and acceptability. Definitions of the psychometric properties and scale quality criteria assessed are summarized in Table I. Various different statistical methods are used to assess the psychometric properties of outcome measurements.8-11 Before the literature search and in accordance with the literature,9 we defined criteria for “adequate” and “acceptable” psychometric properties (Table I).

Criterion validity is the extent to which a measurement relates to an external (gold) standard. Construct validity is given when a scale measures the construct it aims to measure (ie, disease severity). One of the reasons why we did this study is that there currently is no gold standard to assess objective disease severity in AE. Therefore we were unable to consider the aspect of criterion validity in this review. Instead, we tested, for both aspects, construct validity with correlations between 2 different measurements of disease severity relating to convergent validity and correlations between a measurement of disease severity and a measurement of another construct (eg, quality of life) relating to divergent validity (Table I).

Assessment of content validity of domains and items included in outcome measurements

Because we did not find sufficient data to judge content validity, we assessed this issue in a separate study. We selected 12 consumers (from each Department of Dermatology involved in this review [Nottingham, United Kingdom and Dresden, Germany]; 2 patients aged ≥18 years, 2 patients aged 8-14 years, and 2 caregivers of patients aged 1-7 years) and 6 dermatology experts who were not involved in scale development.

Experts and consumers rated content validity of all domains (eg, intensity of lesion and extent of disease) and items (eg, erythema, papulation, and scaling) included in named outcome measurements of AE on a 5-point Likert scale (‘‘very important,’’ ‘‘important,’’ ‘‘indifferent,’’ ‘‘may not be important,’’ and ‘‘unimportant’’). Consumers and experts were blinded to the name of the outcome measurement and assessed the individual domains and items without knowing to which measurement or measurements they belonged. A median10 rating of ‘‘important’’ or ‘‘very important’’ was required to rate a domain or item as adequate. More than 50% of the items used a particular outcome measurement to describe a domain needed to be rated as ‘‘important’’ or ‘‘very important’’ to conclude that the domain was measured adequately.

Criteria applied for recommendation of outcome measurements

Our recommendation on whether to apply an outcome measure is based on the following 8 characteristics: content validity by expert, content validity by consumer, convergent construct validity, divergent construct validity, internal consistency, interobserver reliability, test-retest reliability, and sensitivity to change.

Full credit (100%) was given for characteristics “adequately met,” half credit (50%) for those “acceptably met,” and no credit (0%) for those “not acceptably met,” not assessed, or both (Table I). We used weighted mean ratings (weighted by the number of study
In this study, we considered characteristics that have been assessed in more than one study. For each outcome measurement, we calculated a total relative score ranging from 100%, indicating that all criteria were adequately met, to 0%, indicating that none of the criteria were acceptably met. To grade our recommendation on whether to apply an outcome measurement, we used a 5-point Likert scale with adjectival descriptors.

### RESULTS

We identified a total of 45 eligible articles, 21 of which were retrieved by searching MEDLINE and EMBASE, 17 by hand-searching reference lists, and 6 by free internet search; 1 article was provided by a contacted person. The 45 articles reported on 20 different objective outcome measures for severity of AE. Fig E1 (available at www.jacionline.org) provides additional information on the study identification and selection process. Most validation studies (defined as a study in which at least 1 psychometric property is assessed, Table I) were performed on the Severity Scoring of Atopic Dermatitis index (SCORAD, \( n = 14 \)), Ecema Area and Severity Index (EASI, \( n = 5 \)), and Nottingham Eczema Severity Score (NESS, \( n = 5 \)). We did not identify

<table>
<thead>
<tr>
<th>Name of quality item</th>
<th>Definition of quality item</th>
<th>Measurement of quality item</th>
<th>Criteria for rating “adequate”</th>
<th>Criteria for rating “acceptable”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Construct validity:</td>
<td>Does the scale measure the hypothetical construct (objective severity of AE) it should?</td>
<td>(a) Are 2 outcome measurements that are presumed to measure the same latent construct correlated?</td>
<td>(a) Factor loading/correlation coefficient &gt; 0.70</td>
<td>(a) Factor loading/correlation coefficient 0.60-0.69</td>
</tr>
<tr>
<td>(a) convergent</td>
<td></td>
<td>(a) and (b) Confirmatory factor analysis, Structural equations modeling (correlation of coefficients)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(b) divergent</td>
<td></td>
<td>(b) factor loading/correlation coefficient &lt; 0.70</td>
<td>(b) Factor loading/correlation coefficient 0.71-0.85</td>
<td></td>
</tr>
<tr>
<td>Content validity</td>
<td>Are the domains adequate to measure the construct in question? Are the items representative of the domain they are supposed to measure?</td>
<td>Rating by experts and consumers</td>
<td>Expert/consumer says yes for at least 90% of all items</td>
<td>Expert/consumer says yes for 70% to 89% of all items</td>
</tr>
<tr>
<td>Internal consistency</td>
<td>Are the different domains/items of the scale interrelated?</td>
<td>Cronbach ( \alpha ) *</td>
<td>( \geq 0.90 ) (individual patients)</td>
<td>0.70-0.89 (individual patients)</td>
</tr>
<tr>
<td>Interobserver reliability</td>
<td>Do 2 or more independent investigators achieve the same result?</td>
<td>(a) Correlation coefficient</td>
<td>( &gt; 0.80 )</td>
<td>(a) 0.60-0.80</td>
</tr>
<tr>
<td>Test-retest reliability</td>
<td>Do 2 assessments by one investigator in the same patient yield the same result?</td>
<td>(b) ( \kappa ) †</td>
<td>( &gt; 0.60 )</td>
<td>(b) 0.41-0.60</td>
</tr>
<tr>
<td>Sensitivity to change</td>
<td>Can clinically relevant changes be detected by this measurement?</td>
<td>(c) Coefficient of variation</td>
<td>( &lt; 20% )</td>
<td>(c) 20% to 30%</td>
</tr>
<tr>
<td>Acceptability</td>
<td>Is the measurement practical enough to be applied in:</td>
<td>(d) ANOVA (% variance explained by observer)</td>
<td>( &lt; 10% )</td>
<td>(d) 10% to 20%</td>
</tr>
<tr>
<td></td>
<td>(a) everyday clinical practice</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(b) clinical trials</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Time to administer</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*The Cronbach \( \alpha \) value assesses the extent to which the items and domains of an outcome can be treated as measuring a single latent variable (range, 0 to 1.0; higher values reflect better internal consistency). 9
†The \( \kappa \) value is the chance-corrected agreement between 2 observers (range, 0 to 1.0; higher values reflect higher interobserver reliability). 10

References 23, 24, 31, 32, 35, 44-51, and 56.
any data on the validity of 5 published outcome measurements (Atopic Dermatitis Severity Index, 17 Four Step Severity Score [FSSS], 26 Skin Intensity Score, 52 Six-area Total Body Severity Assessment, 54 and atopic dermatitis severity score [WAZ-S] 55). Table II 13-56 summarizes information of the settings and study populations in validation studies by outcome measurement. Most outcome measurements (Atopic Dermatitis Assessment Measure, 13 Atopic Dermatitis Area and Severity Index, 14, 15 Basic Clinical Scoring System, 18 EASI, 20 Investigators’ Global Atopic Dermatitis Assessment [IGADA], 27 Leicester index, 28 Objective Severity Assessment of Atopic Dermatitis [OSAAD], 34 Rajka and Lange-lind score [RL score], 37 Self-administered Eczema Area and Severity Index [SA-EASI], 39 and Simple Scoring System [SSS] 53) were exclusively validated in secondary or tertiary care settings. For both EASI and SCORAD, the study populations consisted of more than 1000 patients and included infants, children, and adults.

**Domains and items of outcome measurements for AE**

Table III § summarizes the domains of all outcome measurements identified, relative weights of domains in the summary score, and items used to assess the domains. In the 20 outcome measurements, a total of 5 distinct domains were identified: intensity of lesion, extent of disease/body sites affected, symptoms, course of disease, and epidermal function. Substantial heterogeneity between the outcome measurements was found with regard to the domains being included in the summary score, the items used to measure the domains, the relative weights of the domains, the scales used to measure the items, and the person performing the assessment. Most outcomes (n = 18) § include 2 or 3 domains, whereas 2 scales 18, 46 are unidimensional. Disease intensity is assessed in 17 outcome measurements § by a total of 13 different items. In most outcome measurements, physicians are asked to

---

**TABLE II. Characteristics of validation studies on outcome measurements included**

<table>
<thead>
<tr>
<th>Outcome measurement</th>
<th>No. of validation studies</th>
<th>Geographic location</th>
<th>Setting</th>
<th>Study population</th>
<th>No. of participants</th>
<th>Age (range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADAM 13</td>
<td>12, 13</td>
<td>Australia</td>
<td>Secondary/tertiary care</td>
<td>171</td>
<td>0-16 y</td>
<td></td>
</tr>
<tr>
<td>ADASI 14, 15</td>
<td>16</td>
<td>Germany</td>
<td>Secondary/tertiary care</td>
<td>16</td>
<td>1-34 y</td>
<td></td>
</tr>
<tr>
<td>ADSI 17</td>
<td>0</td>
<td></td>
<td>Secondary/tertiary care</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BCSS 18</td>
<td>19</td>
<td>The Netherlands</td>
<td>Secondary/tertiary care</td>
<td>82</td>
<td>0-67 y</td>
<td></td>
</tr>
<tr>
<td>EASI 20</td>
<td>21-25</td>
<td>Australia, United States, Europe, South America</td>
<td>All secondary/tertiary care</td>
<td>1801</td>
<td>0-43 y</td>
<td></td>
</tr>
<tr>
<td>FSSS 26</td>
<td>0</td>
<td>Australia</td>
<td>Secondary/tertiary care</td>
<td>1751</td>
<td>0-17 y</td>
<td></td>
</tr>
<tr>
<td>IGADA 27</td>
<td>22, 24</td>
<td>Australia, United States, South America</td>
<td>Secondary/tertiary care</td>
<td>123</td>
<td>0-60 y</td>
<td></td>
</tr>
<tr>
<td>Leicester index 28</td>
<td>128</td>
<td>United Kingdom</td>
<td>Community, n = 2; secondary/tertiary care, n = 3</td>
<td>651</td>
<td>1-63 y</td>
<td></td>
</tr>
<tr>
<td>NESS 29</td>
<td>20-33</td>
<td>United Kingdom, China</td>
<td>Secondary/tertiary care</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OSAAD 34</td>
<td>34, 35</td>
<td>Europe, United States</td>
<td>All secondary/tertiary care</td>
<td>70</td>
<td>0-38 y</td>
<td></td>
</tr>
<tr>
<td>POEM 36</td>
<td>16</td>
<td>United Kingdom</td>
<td>Primary and secondary care</td>
<td>453</td>
<td>1-62 y</td>
<td></td>
</tr>
<tr>
<td>RL score 37</td>
<td>16, 25, 38</td>
<td>Europe</td>
<td>All secondary/tertiary care</td>
<td>52</td>
<td>1-54 y</td>
<td></td>
</tr>
<tr>
<td>SA-EASI 39</td>
<td>39, 40</td>
<td>United States</td>
<td>All secondary/tertiary care</td>
<td>96</td>
<td>0-12 y</td>
<td></td>
</tr>
<tr>
<td>SASSAD 41</td>
<td>33, 42, 43</td>
<td>United Kingdom, Australia</td>
<td>Primary, n = 1; secondary/tertiary, n = 2</td>
<td>124</td>
<td>3-63 y</td>
<td></td>
</tr>
<tr>
<td>SCORAD 44</td>
<td>14, 23, 24, 31, 32, 35, 44, 51, 56</td>
<td>Europe, China, Canada</td>
<td>Community, n = 4; secondary/tertiary care, n = 11</td>
<td>1346</td>
<td>0-67 y</td>
<td></td>
</tr>
<tr>
<td>SIS 52</td>
<td>0</td>
<td></td>
<td>Secondary/tertiary care</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSS 53</td>
<td>46, 19, 28, 53</td>
<td>Europe</td>
<td>All secondary/tertiary care</td>
<td>235</td>
<td>0-67 y</td>
<td></td>
</tr>
<tr>
<td>TBSA 54</td>
<td>0</td>
<td></td>
<td>Primary and secondary care</td>
<td>306</td>
<td>0-67 y</td>
<td></td>
</tr>
<tr>
<td>TISS 46</td>
<td>46, 51</td>
<td>Europe</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WAZ-S 55</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ADAM, Atopic Dermatitis Assessment Measure; ADASI, Atopic Dermatitis Area and Severity Index; NA, not applicable; ADSI, Atopic Dermatitis Severity Index; BCSS, Basic Clinical Scoring System; SIS, Skin Intensity Score; TBSA, Six-area Total Body Severity Assessment.

---

†References 13, 15, 17, 18, 20, 26-29, 34, 36, 37, 39, 41, 44, 46, and 52-55.
‡References 13, 15, 17, 20, 26-29, 34, 36, 37, 39, 41, 44, and 52-55.
§References 13, 15, 17, 20, 26-28, 34, 36, 37, 39, 41, 44, 46, and 52-55.
grade intensity on Likert scales, whereas others (eg, SA-
EASI39) use visual analog scales marked by the patient or
caregiver. Disease intensity attributes 33%15 to
100%46 to the summary score. An assessment of disease
extent is included in 16 outcome measurements** and
attributes 19%;44 to 100%;18 of the summary score. An
estimation of the involved body surface area is required by
9†† and involvement of special body sites by
7;15,27,28,41,53,55 outcome measurements assessing disease
extent. Body surface area is measured heterogeneously,
with some outcome measurements applying the “rule of
nines” (eg, SCORAD44), others using tick boxes (eg,
NESS29), and others using a silhouette on which the
patient marks involved body sites (eg, SA-EASI19).
Disease symptoms like pruritus are assessed in 11 out-
come measurements†† attributing up to 33% to the total
score (Table III). FSSS,26 NESS,29 and RL score37 include
an assessment of the course of disease within the past year.
Epidermal function is measured exclusively in the
OSAAD.34

Assessment of content validity of domains and
items included in outcome measurements

Both experts and consumers considered intensity of
lesions and extent of disease as “very important” criteria
for the assessment of the severity of AE. Course of disease
and symptoms were also judged to be “important” or
“very important” by both groups, whereas epidermal
function was considered as “may not be important” by the
experts and “indifferent” by the consumers.

Experts tended to rate items that are less specific
for AE as less important when assessing disease severity.
Although consumers considered cracking/fissuring, ves-
cles, bleeding, and erosions as “very important,” experts
were “indifferent” or judged these items as “may not be
important” (Fig 1).

Validity of outcome measurements and
recommendation

Table IV§§ outlines the results of validation studies on
all outcome measurements identified. Content validity, as
assessed by the consumer, is adequate for all outcomes
except the OSAAD.34 Based on the experts’ rating,
OSAAD,34 Patient-oriented Eczema Measure [POEM],36
and WAZ-S55 do not have acceptable content validity.
Only EASI,22-24 SCORAD,23,24,34 and the Three Item
Severity Score (TISS)36,51 have been shown to have ade-
quate convergent and divergent construct validity.

Evidence for adequate internal consistency was found
only for the POEM.36 Eighteen outcome measurements
had either unacceptable internal consistency (n = 4;

**References 13, 15, 18, 20, 26-29, 34, 37, 39, 41, 44, and 53-55.
††References 15, 18, 20, 26, 29, 34, 37, 39, and 44.
†††References 13, 15, 17, 29, 36, 37, 39, 44, 52, 53, and 55.
§§References 12, 12-22, 24, 26-46, and 49-56.
***References 13, 17, 18, 26-29, 34, 39, 41, 46, 52, 54, and 55.
††††References 13-15, 17, 18, 26, 29, 46, 52, 54, and 55.

Atopic Dermatitis Area and Severity Index,14-16
SSS,16,53 SCORAD,39,50 and RL score16,37) or have not
been validated for internal consistency (n = 14).***
There is convincing evidence to conclude that BSCC15,9,
NESS,22; OSAAD;34 Six Area, Six Sign Atopic
Dermatitis severity score [SASSAD];25, and
SCORAD,35,44,46 have adequate interobserver reliability,
whereas adequate test-retest reliability has been shown
only for the POEM.36 For most of the outcome measure-
ments, identified interobserver reliability and test-retest
reliability have not been evaluated adequately yet.
Sensitivity to change is adequate for EASI,24 IGADA,24
and SCORAD24; acceptable for Leicester index,28
OSAAD,35 POEM,36 SA-EASI,40 SASSAD,33 and
SSS;28; not acceptable for the RL score25; and has not
been adequately assessed for the remaining 10 mea-
surements..†††† The time needed to perform disease severity
assessment ranges from 1 minute16,36 up to 10
minutes (Table IV).16,41,44,56

Based on the existing evidence, none of the 20 outcome
measurements can be highly recommended. EASI,
POEM, and SCORAD have been shown to meet most
validity criteria and are recommended for use. A detailed
description of EASI, POEM, and SCORAD is given in
the online version of this article (see the online Results
section available in the Online Repository at www.jacionline.
org). Although the validity criteria are only partly met,
IGADA, NESS, SA-EASI, SASSAD, and TISS appear
to be acceptable until further validation studies are avail-
able. Because of a lack of evidence for their validity, we
do not recommend using the remaining 12 outcome
measurements (Table IV).

There was no association between the time since
inauguration of outcome measurements (regression coef-
ficient, −0.01; P = .99) and their validity.

DISCUSSION

Main findings

Currently, investigators can select from 20 different
named measurements of disease severity. Of these, only
EASI, POEM, and SCORAD have been validated ade-
quately enough at present to recommend their use in
clinical trials and everyday practice.

Our study extends previous research because it applies
objective criteria to judge whether the psychometric
properties of existing outcome measurements are ade-
quate. The reason why we currently do not recommend
using 17 of the 20 outcome measurements identified is
primarily that data on their validity is missing rather than
an indication of inappropriateness of the instruments
because existing data show that they are not valid. Since
the review by Charman and Williams in 2000,7 7 new out-
come measurements have been introduced (FSSS,26
IGADA,22 OSAAD,34 POEM,36 SA-EASI,39 TISS,36
and WAZ-S55). Of these, however, only the POEM has
been adequately validated (Table IV).

Most outcome measurements analyzed are assessed by
a physician (eg, EASI and SASSAD) and are therefore
sometimes referred to as “objective,” whereas others (eg, POEM, SA-EASI) are more “subjective” because they are scored directly by the patient or caregiver. In reality, of course, both measurements require subjective judgment of categories within domains by physician and patient. The disadvantage of “subjective” outcome measurements is that bias might be introduced when coping strategies, quality-of-life impairment, or comorbidity influence rating. The advantage of a subjective measurement like the POEM, however, is that it truly measures what is important to the patient. Although some of the items included in the POEM were not considered as adequate by our experts, the POEM was shown to be highly valid from the consumers’ perspective (Table IV).

### Study strengths and limitations

Based on objective criteria concerning their psychometric properties, we made recommendations on which outcome measurements to apply. Our recommendations were informed by a systematic and comprehensive literature search. The cutoffs used to judge whether the validity criteria studied are “adequate” or “acceptable” are consistent with the literature and were defined a priori (ie, before an independent group looked at the data).6,10

Despite our efforts to come up with evidence-based and objective recommendations, we know that recommendations are always to some extent subjective and therefore subject to discussion.

Another potential limitation of this study is that we did not consider acceptability (ie, time needed to perform measurement) in our recommendation. We decided to do so because the amount of time needed highly depends on the experience of the person doing the assessment and because it was not clear from most of the articles whether the time needed was really measured rather than just estimated.

By including content validity from both a consumer’s and expert clinician’s perspective, content validity has been included twice in our overall grading system, and we acknowledge that such an approach might have given disproportionate weight to such a domain. We believe that consumer and expert perspectives are measuring slightly different things. Combining both content validity perspectives into one overall composite score did not alter the overall conclusions or choice of instruments in our review (data not shown).

To assess content validity, we surveyed different consumer groups, which raises the question of whether the responses were homogeneous across those groups. We found that median responses of adult patients, children aged 8 to 14 years, and parents of younger patients were almost identical (data not shown). This finding provides good evidence for the validity of our approach to assess content validity.

Some authors have used the Investigators’ Global Assessment (IGA) as a gold standard and assessed criterion validity of other outcome measurements by correlating...
them with the IGA.\textsuperscript{7} We believed that the IGA is probably not stable enough to be a gold standard because factors like response to previous treatments, compliance, and aspects of the patient-physician relationship contribute to a global assessment of disease severity. The IGA has never been inaugurated as an outcome measurement, and we are not sure when it was introduced. The IGADA is a variant of the IGA in which objective rules on how to rate severity are defined.\textsuperscript{27} The IGADA gives verbal descriptions for disease severity, such as “almost clear” or “very severe,” which seems to be useful for clinical practice. Future research is necessary to evaluate the reliability of the IGADA.

### Implications for current research and clinical practice

We found substantial heterogeneity in the domains included in the different outcomes, the items used to measure the domains, the relative weights of the domains on the summary score, the scales used to measure the items, and the person performing the assessment. This leads to the conclusion that the 20 named outcomes identified do not measure the same thing. We recommend using EASI or (objective) SCORAD as a valid and unbiased estimate of “objective” disease severity plus the POEM as a measurement of AE severity from the patient’s perspective.

### Implications for future research

Our study highlights that validation and standardization of the outcomes methodology in clinical research...
are key issues for future research. We plan to use the results of this review to undertake an international consensus project involving both clinical experts and epidemiologists on a core set of end points for RCTs, effectiveness studies, and clinical record keeping in daily practice. Preferences achieved with a high degree of consensus might help to establish an international standard for clinical investigations.

We thank our patients and Dr Paula E. Beattie (Consultant Dermatologist, Western Infirmary, Glasgow, United Kingdom), Dr Sue Lewis-Jones (Consultant Dermatologist, Department of Dermatology, Ninewells Hospital and Medical School, Dundee, Scotland), Dr Ruth Murphy (Consultant Dermatologist, Nottingham University Hospitals NHS Trust, United Kingdom), Dr Amy S. Paller (Professor and Chair, Department of Dermatology, Feinberg School of Medicine, Northwestern University, Chicago, Ill), Dr Jane Ravenscroft (Department of Dermatology, Nottingham University Hospitals NHS Trust, United Kingdom), and Dr Torsten Schäfer (Professor, Department of Social Medicine, Lübeck, Germany) for their support in the assessment of content validity. We also thank Dr Finola Delamere, trials search coordinator of the Cochrane Skin Group, for her assistance in the electronic literature search.

REFERENCES


### TABLE IV. (Continued)

#### Psychometric properties and scale quality

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Individual studies</th>
<th>Total†</th>
<th>Individual studies</th>
<th>Total†</th>
<th>Individual studies</th>
<th>Total†</th>
<th>Time needed to perform (min)</th>
<th>Mean Score</th>
<th>Recommendation*</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADAM††</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>44%</td>
<td>–</td>
</tr>
<tr>
<td>ADASI15</td>
<td>O13</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>2-10</td>
<td>38%</td>
<td>–</td>
</tr>
<tr>
<td>ADSI17</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>25%</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>BCSS18</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>44%</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>EASI20</td>
<td>O21</td>
<td>0</td>
<td>O21</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>81%</td>
<td>–</td>
<td>+</td>
</tr>
<tr>
<td>FSSS16</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>25%</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>IGADA27</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>50%</td>
<td>–</td>
<td>+</td>
</tr>
<tr>
<td>Leicester index28</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>31%</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>NESS29</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>1-32</td>
<td>56%</td>
<td>+</td>
</tr>
<tr>
<td>OSAAD34</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>5-34</td>
<td>25%</td>
<td>–</td>
</tr>
<tr>
<td>POEM16</td>
<td>A1</td>
<td>0</td>
<td>A1</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>1-4-16</td>
<td>31%</td>
<td>–</td>
</tr>
<tr>
<td>RL score37</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>1-2-16</td>
<td>71%</td>
<td>+</td>
</tr>
<tr>
<td>SA-EASI39</td>
<td>N4</td>
<td>0</td>
<td>N4</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>1-4-16</td>
<td>57%</td>
<td>–</td>
</tr>
<tr>
<td>SASSAD41</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>&lt;2-10-41</td>
<td>56%</td>
<td>–</td>
</tr>
<tr>
<td>SCORAD44</td>
<td>O46</td>
<td>0</td>
<td>O46</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>&lt;10-45-56</td>
<td>81%</td>
<td>+</td>
</tr>
<tr>
<td>SIS52</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>25%</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>SSS53</td>
<td>O19</td>
<td>0</td>
<td>O16</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>1-5-16</td>
<td>44%</td>
<td>–</td>
</tr>
<tr>
<td>TBSA54</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>25%</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>TISS56</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>56%</td>
<td>–</td>
<td>+</td>
</tr>
<tr>
<td>WAZ55</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>13%</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

---

学霸图书馆（www.xuebalib.com）是一个“整合众多图书馆数据库资源，提供一站式文献检索和下载服务”的24小时在线不限IP图书馆。
图书馆致力于便利、促进学习与科研，提供最强文献下载服务。

图书馆导航：
图书馆首页 文献云下载 图书馆入口 外文数据库大全 疑难文献辅助工具