

Received Date : 25-Mar-2016

Revised Date : 21-May-2016

Accepted Date : 18-Jun-2016

Article type : Original Article: Skin and Eye Diseases

Evaluation of the measurement properties of symptom measurement instruments for atopic eczema: a systematic review

Louise A.A. Gerbens, MD,¹ Cecilia A.C. Prinsen, PhD,² Joanne R. Chalmers, PhD,³ Aaron M. Drucker, MD,⁴ Laura B. von Kobyletzki, MD, MSc, PhD,⁵ Jacqueline Limpens, PhD,⁶ Helen Nankervis, PhD,³ Åke Svensson, MD, PhD,⁵ Caroline B. Terwee, PhD,² Jianzhong Zhang, MD, PhD,⁷ Christian J. Apfelbacher, MSc, PhD,⁸ Phyllis I. Spuls, MD, PhD,¹ on the behalf of the Harmonising Outcome Measures for Eczema (HOME) initiative.

¹ Department of Dermatology, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands

² Department of Epidemiology and Biostatistics, EMGO+ Institute for Health and Care Research, VU University Medical Centre, Amsterdam, The Netherlands

³ Centre of Evidence Based Dermatology, University of Nottingham, Nottingham, UK

⁴ Division of Dermatology, University Health Network, Toronto, Canada and Department of Dermatology, Brown University, Providence, RI, USA

⁵ Institution for Clinical Science, Department of Dermatology, Lund University, Malmö, Sweden

⁶ Medical Library, Academic Medical Centre, Amsterdam, the Netherlands

⁷ Department of Dermatology, Peking University People's Hospital, Peking, China

⁸ Medical Sociology, Institute of Epidemiology and Preventive Medicine, University of Regensburg, Regensburg, Germany

Corresponding author:

Louise A.A. Gerbens

Department of Dermatology, Academic Medical Centre, University of Amsterdam, The Netherlands
Meibergdreef 9

1105 AZ Amsterdam, The Netherlands

Phone: +31 20 5662581

Fax: + 31 20 5669152

E-mail: l.a.gerbens@amc.uva.nl

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/all.12959

This article is protected by copyright. All rights reserved.

Abstract

Background

Symptoms have been identified as a core outcome domain for atopic eczema (AE) trials. Various instruments exist to measure symptoms in AE, but they vary in quality and there is a lack of standardisation between clinical trials.

Our objective was to systematically evaluate the quality of the evidence on the measurement properties of AE symptom instruments, thereby informing consensus discussions within the Harmonising Outcome Measures for Eczema (HOME) initiative regarding the most appropriate instruments for the symptoms core outcome domain.

Methods

By using the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) checklist and predefined criteria for good measurement properties on identified development and validation studies of AE symptom instruments, a best evidence synthesis was performed to draw an overall conclusion on quality of the instruments and to provide recommendations.

Results

Eighteen instruments were identified and evaluated. When the quality and results of the studies were considered, only five of these instruments had sufficient validation data to consider them for the core outcome set for the core outcome domain symptoms. These were the paediatric Itch Severity Scale (ISS), Patient-Oriented Eczema Measure (POEM), Patient-Oriented SCOring Atopic Dermatitis (PO-SCORAD), Self-administered Eczema Area and Severity Index (SA-EASI) and adapted SA-EASI.

Conclusions

ISS (paediatric version), POEM, PO-SCORAD, SA-EASI and adapted SA-EASI are currently the most appropriate instruments and therefore have the potential to be recommended as core symptom instrument in future clinical trials. These findings will be utilised for the development of a core outcome set for AE.

Background

Various interventions are available for the treatment of atopic eczema (AE) (also known as atopic dermatitis or eczema), one of the most common dermatologic conditions with a lifetime prevalence of up to 10% in adults (1). There is heterogeneity and lack of standardisation of outcome measurement instruments (further referred to as 'instruments') to measure the outcomes (i.e. constructs or domains) of interventions in AE trials (2). This causes difficulties in comparing and pooling outcomes and, as a consequence, hampers evidence-based clinical decision making.

The international Harmonising Outcome Measures for Eczema (HOME) initiative aims to develop core outcome sets (COS) for four previously agreed upon core outcome domains for AE: clinical signs, symptoms, health-related quality of life, and long-term control (3). COS are agreed minimum sets of outcomes that should be measured and reported in all clinical trials of a specific disease or trial population (4). The HOME roadmap serves as a guide to develop a COS for AE, which involves

subsequent steps and stages that lead to a recommendation of an instrument (5).

In this systematic review we aim to evaluate the evidence on the measurement properties of instruments for the core outcome 'symptoms', defined as a departure from normal function, appearance or feeling which is noticed by the patient, indicating the presence of disease or abnormality (6). Examples of symptoms are itch, pain and redness. In contrast, clinical signs are noticed by the physician and thus measured using a physician-assessed instrument. A recent systematic review completed Stage 1 of the HOME roadmap, which identified the instruments used to measure symptoms in AE treatment randomised controlled trials (7).

The specific objectives of the present systematic review were:

- (1) to systematically appraise the measurement properties of instruments for AE symptoms;
- (2) to categorise the instruments according to predefined criteria, as instruments
 - (a) that meet the predefined criteria to be recommended for the measurement of symptoms in future AE trials,
 - (b) that have the potential to be recommended in the future depending on the results of further validation studies,
 - (c) that do not meet the predefined criteria to be recommended and therefore should not be used anymore;
- (3) to provide the evidence base for the HOME international consensus process to further standardize the symptom instruments in clinical trials and to prioritize further research concerning AE symptom instruments.

Methods

This systematic review was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) statement (8). The study protocol has been registered in the International Prospective Register of Systematic Reviews (PROSPERO): CRD42015020295.

Literature search

A systematic literature search was carried out by a medical librarian [JL] in OVID MEDLINE, OVID EMBASE and the publisher-subset of PubMed on August 27, 2014 with an update on August 3, 2015. The main search strategy consisted of three blocks of search terms: 1) target population (AE); 2) construct of interest (symptoms); and 3) a validated highly sensitive search filter for measurement properties (9), that was adapted for OVID MEDLINE and EMBASE and supplemented with terms for interpretability and feasibility. For completeness, a title search for AE instruments was also performed (without applying any search filter). No language or other restrictions were applied. The full MEDLINE search strategy can be found in the Supplemental Material Table 1. Reference lists of papers included were screened for potentially eligible studies.

Eligibility criteria

Studies were selected if published as full-text papers and if their purpose was to develop an instrument to measure symptoms in AE ('development or inauguration paper') and/or evaluate the measurement properties of an instrument that measures symptoms in patients with AE ('validation paper').

Eligible instruments were those which were specifically developed and/or validated to measure symptoms in children and adults with AE to assess disease severity, and also any composite instrument that included symptoms and which separately reported on the validation of the symptom items. Instruments that had not been validated were not eligible. A complete overview of the inclusion and exclusion criteria can be found in Supplemental Material Table 2.

Study selection

Two reviewers [LG/PS] independently screened titles and abstracts, to be followed by the assessment of full texts for eligibility (in Supplemental Material Table 2). A third reviewer [CP] was consulted in case of uncertainties or disagreements.

Data extraction

Relevant data from all included papers were extracted using a study-specific and pilot-tested evidence table. The characteristics of included studies and instruments, results on measurement properties (reliability, validity, responsiveness), and information on interpretability and feasibility were extracted. Where information could not be obtained through the included papers, the original development paper was consulted. Data were extracted independently by three trained review teams who worked in pairs on defined sets of papers [LG/AD, JC/HN, and PS/JZ]. Disagreements were resolved by discussion within the research team.

Assessment of the methodological quality of included studies

To evaluate the methodological quality of the included studies, the CONsensus-based Standards for the selection of health Measurement INstruments (COSMIN) checklist with a four-point rating scale was used (10). Assessment was performed independently by three trained review teams who worked in pairs [LG/CP, LvK/AS, and CA/PS]. Disagreements were resolved by means of discussion within the research team. Where a single paper presented multiple studies, each study was assessed and rated separately.

Evaluation of the study results against criteria for good measurement properties

We used the COSMIN taxonomy (Supplemental Material Table 3) to decide which measurement properties of an instrument were evaluated, and the results of the studies were evaluated against predefined criteria for good measurement properties (Supplemental Material Table 4) (11). As a gold standard for 'symptoms' is lacking, 'criterion validity' was not considered.

Rating the quality of the instrument by performing a best evidence synthesis

Overall evidence on each measurement property for each instrument was synthesized using the results on measurement properties of the different studies, adjusted for the methodological quality of the studies, by applying the levels of evidence (Table 1) (11, 12). Studies on the same instrument performed with different age groups were combined if no separate validation data were present as were different language versions of the same instrument.

Generating recommendations for the use of AE symptom outcome measurement instruments

Based on the results of the best evidence synthesis, a standardized recommendation for usage or future validation work was made for each instrument identified in the review, reflecting the three criteria of the Outcome Measures in Rheumatology (OMERACT) filter: truth, discrimination, and feasibility (13, 14).

Recommendations were defined as follows: category A, symptom instrument meets all requirements (positive rating for all boxes in the best evidence synthesis) and is recommended for use; B, symptom instrument meets two or more required quality items, but performance in all other required quality items is unclear, so that the instrument has the potential to be recommended in the future depending on the results of further validation studies; C, symptom instrument has low quality in at least one required quality criteria (≥ 1 rating of 'minus') and therefore is not recommended to be used anymore; D, symptom instrument has almost not been validated. Its performance in all or most relevant quality items is unclear, so that it is not recommended to be used until further validation studies clarify its quality.

Results

Of the 2332 papers identified by our search, 29 eligible papers were included in this review (Supplemental Material Figure 1), containing 40 studies evaluating 18 different instruments. Identified instruments in papers that were excluded, with reasons, are summarised in the online appendix (Supplemental Material Table 5). In several studies ($n=4$), more than one instrument was investigated. Table 2 shows the characteristics of the included studies. Table 3 presents the characteristics of the included instruments.

Quality assessment and results of the studies

The methodological quality of the included studies and the evaluation of the results are presented in Table 4 and 5 respectively.

Quality of the instruments: best evidence synthesis

The results of the best evidence syntheses (Table 1 and 6) are described below per instrument according to the category of recommendation (A-D).

Category A instruments

Currently no instruments meet all requirements to be recommended for the measurement of symptoms in future AE trials and to be included in the COS for AE.

Category B instruments

Five instruments met the requirements for a category B recommendation and have the potential to be recommended for use in clinical trials and to be included in the COS for AE. Their pros and cons are summarised in Supplemental Material Table 6.

Paediatric Itch Severity Scale (ISS)

One study evaluated the measurement properties of paediatric Spanish ISS (15). Limited evidence was found for good internal consistency, test-retest reliability, structural validity and convergent validity. There was limited evidence for indeterminate responsiveness. Interpretability yielded a minimal important difference (MID) of 3 points.

Patient-Oriented Eczema Measure (POEM)

The measurement properties of the POEM were assessed in seven studies (16-19). Limited evidence was found for good internal consistency (16), moderate evidence for good construct validity (16, 19) and for good responsiveness (16, 17, 19). Due to poor methodological study quality the test-retest reliability remains unknown (16). The evidence for measurement error could not be interpreted because of one study with limited evidence for indeterminate measurement error (16) and one study of poor methodological quality (17). Good content validity was supported by moderate evidence (16). Interpretation was assessed and yielded a minimal clinically important difference (MCID) of 3.4 points (17), five bands of severity (i.e. clear, mild, moderate, severe, very severe) (18), and a mean absolute change in score from baseline of 7.9 (standard deviation (SD) 6.0) (19).

Patient-Oriented SCORing Atopic Dermatitis (PO-SCORAD)

Four studies evaluated the measurement properties of the PO-SCORAD (19-21). Limited evidence was found for good content validity (20) and responsiveness (19). There was conflicting evidence for construct validity (19-21). Interpretability was assessed and showed a mean absolute change in score from baseline of 24.5 (SD 17.1) (19).

Self-Administered Eczema Area and Severity Index (SA-EASI)

Three studies addressed the measurement properties of the SA-EASI (22-24). Limited evidence was available for good inter-rater reliability of the Italian SA-EASI (23). Since one study of poor methodological quality assessed content validity of the English SA-EASI, this evidence remains unknown (22). There is limited evidence for good construct validity of the English SA-EASI (22, 24) besides unknown evidence due to poor study quality for the Italian SA-EASI (23).

Adapted Self-Administered Eczema Area and Severity Index (adapted SA-EASI)

One study investigated the measurement properties of the adapted version of the SA-EASI (19). There was limited evidence for good construct validity and good responsiveness. Interpretability demonstrated a mean absolute change in score from baseline of 4.9 (SD 5.4).

Category C instruments

Details of category C instruments are described in the Supplementary Material. Six instruments had a low quality in at least one required quality criterion (i.e. measurement property) and are recommended not to be used in future trials, i.e. Atopic Dermatitis Assessment Measure (ADAM) (25, 26), Eppendorf Itch Questionnaire (EIQ) (27), adult Itch Severity Scale (ISS) (15, 28), Leuven Itch Scale (LIS) (29), Skin Detective Questionnaire (SDQ) (30), Zheng-Related Atopic Dermatitis Symptom Questionnaire (ZRADSQ) (31). The number of measurement properties evaluated ranged from 2 to 6. The ADAM and LIS had low quality in two measurement properties, the others in one. Reliability was poor for ADAM, adult ISS, LIS and SDQ, structural validity was poor for ADAM and ZRADSQ, and hypothesis testing was poor for EIQ and LIS.

Category D instruments

Details of category D instruments are described in the Supplementary Material. Seven instruments had almost not been validated or the results on the measurement properties remain unclear, i.e. Atopic Dermatitis Quickscore (ADQ) (32), web-based Characteristics of Itch Questionnaire (CoIQ) (33), Method 4 (34), Nottingham Eczema Severity Score (NESS) (35-37), subjective SCORAD (27, 32, 36, 38-42), Visual Analogue Scale (VAS) (43) and Verbal Rating Scale (VRS) pruritus (43). These instruments are not recommended to be used until further validation studies clarify their quality. The number of measurement properties evaluated ranged from 1 to 5, with 4 instruments having only 1 property assessed (i.e. ADQ, Method 4, VAS and VRS pruritus). The majority of the measurement properties investigated remained unclear in their evidence. The CoIQ demonstrated limited evidence for good test-retest reliability. Strong evidence was found for good structural validity of the total SCORAD index (Schafer et al., 1997). Principal component analysis demonstrated, however, that the two items of the subjective SCORAD belong to different factors, indicating that part C is not unidimensional (i.e. the items do not measure the same construct). Five studies found evidence for construct validity of the subjective SCORAD, four with conflicting findings (Carel et al., 2008; Darsow et al., 2001; Hon et al., 2006a; Hon et al., 2006b) and unknown findings from one study of poor methodological quality (Pucci et al., 2005).

Discussion

Eighteen instruments for the assessment of symptoms of AE were identified and evaluated in this systematic review. Five instruments are currently the most appropriate available instruments and have the potential to be recommended for the COS based on their known measurement properties: the paediatric ISS, POEM, PO-SCORAD, SA-EASI and adapted SA-EASI.

Although all the identified instruments have gaps in their validation (Table 6) and future validation studies should focus on these, special emphasis for future research should be placed on content validity. Data on comprehensibility of the instrument, relevance of the items for AE patients, for the domain symptoms and the purpose of the measurement, and assurance that no essential items are missing, is of great importance. In this context a survey amongst patients has been conducted to evaluate which symptoms and clinical signs are important to patients in determining treatment effect (44). Itch, pain/soreness, skin feels hot or inflamed, bleeding, involvement of visible or sensitive body sites, cracks, sleep difficulties, amount of body affected and weeping were rated as being quite or very important. For the development of new instruments or validation of existing instruments this survey can be used to evaluate content validity. Both the POEM and PO-SCORAD have demonstrated content validity, whereas for paediatric ISS, adapted SA-EASI and SA-EASI, this evidence is not available or unknown.

Despite our finding that these five instruments are the most appropriate instruments currently available, this is not reflected in clinical trials. In our previous review, it was shown that the POEM was used in only five trials and that the other four instruments (i.e. paediatric ISS, PO-SCORAD, SA-EASI, adapted SA-EASI) were not applied in any trial since 2000 (7).

Our findings provided a basis for standardisation of the symptom instruments in clinical trials. They were used during the HOME IV consensus meeting (Malmö, Sweden, April 23-24, 2015) (6) with the goal of identifying a core instrument to measure symptoms in all future AE clinical trials in order to improve evidence-based decision making in clinical practice. A consensus statement will be published separately.

Search update

The search of this review was updated prior to publication, after the HOME IV meeting. To the original 26 papers three were added (19, 24, 43). Some of the preliminary data presented at HOME IV were re-reviewed. The adapted SA-EASI, VAS and VRS pruritus were added as new instruments. The PO-SCORAD recommendation changed from category D to B due to the inclusion of an extra paper; the subjective SCORAD changed from category C to D; and the SDQ from category D to C due to re-review. Further the ISS was evaluated separately for the adult and paediatric version, which changed the recommendation of the adult ISS from category B to C. The modified EASI (mEASI) was omitted since the content of this instrument remains unclear (15). Additional to the EASI it seems to measure itch, but no published validation paper actually stated its content. Our systematic review results therefore slightly differ from the preliminary results that were used during the meeting. However, it is highly unlikely to have changed the voting results because an essential list of symptoms was composed during the meeting and the content of the instruments was evaluated against this list. Instruments assessing only one symptom were rejected, i.e. ISS, and this argument would apply to the VAS and VRS pruritus. The subjective SCORAD was not considered at the meeting as it is only part of the full SCORAD. However, voting was performed on the PO-SCORAD, and it was not selected as a core instrument. Although more validation data are now available, the argument was raised that this instrument, together with the SA-EASI, is essentially asking patients to perform clinical signs rather than being a true measurement of patient-reported symptoms. For the adapted SA-EASI the same argument applies and, further, validation data are available from only one study.

Strengths and limitations of this systematic review

We followed a registered protocol and adhered to the PRISMA statement (8). Our search was performed using a validated search filter without restrictions, meaning that the search was maximally inclusive. The COSMIN checklist was used to rate the methodological study quality, taking into account possible bias of the results of low quality studies. Reviewers were trained and worked in pairs and disagreements or uncertainties were resolved through discussions within the review team.

Some potential limitations remain. First, three reviewer pairs performed the COSMIN rating, possibly leading to inter-rater variability. Second, studies with different language versions of an instrument and with different age groups were combined. Different language versions may present cross-cultural differences, unless it is proven that no differential item functioning (DIF) for language exists. Since there was no or unknown evidence on cross-cultural validity we have chosen to merge the results of different language versions. The same applies to different age groups; unless no DIF for age exists, paediatric and adult AE patients may warrant specific instruments based on its content. However many validation studies combined children and adults in their study population, and therefore we decided to combine these age groups. Further, the included studies were mainly performed in Western countries and Asia, so results may be different for other locales. Lastly, our review revealed the difficulty of evaluating composite instruments, i.e. instruments measuring symptoms in addition to other domains. Our protocol was designed to be as inclusive as possible and therefore included the composite instruments ADAM and SCORAD index. It could be argued that validation of an instrument should apply to the whole instrument and thus it would be inappropriate to use only one part of the instrument. It could also be argued that a unidimensional subscale could be seen as a separate instrument. To have excluded composite instruments would have meant excluding the SCORAD index. Because SCORAD is so widely used we felt it was appropriate to determine the measurement properties of the subjective domain of this instrument for measuring symptoms.

Recommendations for researchers, clinicians and decision makers

Based on the currently available evidence, the paediatric ISS, POEM, PO-SCORAD, SA-EASI and adapted SA-EASI are currently the most appropriate available instruments and have the potential to be recommended as instruments to assess symptoms of AE in future clinical trials, although future validation research may alter our recommendations.

The deployment of a COS, including the outcome domains of symptoms, clinical signs, quality of life and long-term control, in all future AE clinical trials is a key requirement to improve trial reporting and to facilitate meta-analysis. Stakeholders should report symptoms in AE trials as it has been agreed to be a core outcome domain. Ideally one well-validated core instrument for symptoms and for the other domains will be used in the future; the HOME initiative is the driving force in moving towards this goal.

Abbreviations

COS, Core Outcome Set; COSMIN, COnsensus-based Standards for the selection of health Measurement INstruments; DIF, Differential item functioning; HOME, Harmonising Outcome Measures for Eczema; MCID, Minimal clinically important difference; OMERACT, Outcome Measures in Rheumatology; MID, Minimal important difference (MID); PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analysis; PROSPERO, Prospective Register of Systematic Reviews.

Author contributions

LG wrote the protocol, conceptualized the research plan for the systematic review, screened search results, extracted data, assessed measurement properties, coordinated the work of the other reviewers, wrote the manuscript and reviewed it for important intellectual content. CP reviewed the protocol, screened search results, assessed measurement properties, wrote the manuscript and reviewed it for important intellectual content. JC reviewed the protocol, extracted data and reviewed the manuscript for important intellectual content. AD extracted data and reviewed the manuscript for important intellectual content. LvK reviewed the protocol, assessed measurement properties and reviewed the manuscript for important intellectual content. JL performed the literature search, helped with the methodology section and reviewed the manuscript for important intellectual content. HN extracted data and reviewed the manuscript for important intellectual content. AS reviewed the protocol, assessed measurement properties and reviewed the manuscript for important intellectual content. CT offered methodological support and reviewed the manuscript for important intellectual content. JZ reviewed the protocol, extracted data and reviewed the manuscript for important intellectual content. CA reviewed the protocol, assessed measurement properties and reviewed the manuscript for important intellectual content. PS wrote the protocol, conceptualized the research plan for the systematic review, extracted data, assessed measurement properties, wrote the manuscript and reviewed it for important intellectual content.

Conflict of interest

Aaron Drucker is currently validating an as-yet unpublished novel disease severity assessment tool for AE that is not included in this review. Åke Svensson was involved with the development and validation of the PO-SCORAD. Christian Apfelbacher and Phyllis Spuls are members of the executive committee of the HOME initiative. All authors, except Jacqueline Limpens and Caroline Terwee, are ordinary members of the HOME initiative. The authors state there are no other conflicts of interests.

Acknowledgments

The authors are very grateful for the help of Masutaka Furue (Fukuoka, Japan) for the data extraction and evaluation of the Japanese study, Ellis Hon (Hong Kong, China), Kyu Han Kim (Seoul, South Korea), Takeshi Nakahara (Fukuoka, Japan), Makiko Nakahara (Fukuoka, Japan), Matthew Ridd (Bristol, UK), Jochen Schmitt (Dresden, Germany) and Elke Weisshaar (Heidelberg, Germany) for reviewing the study protocol.

References

1. Bieber T. Atopic dermatitis. *Annals of dermatology*. 2010;22(2):125-37.
2. Charman C, Chambers C, Williams H. Measuring atopic dermatitis severity in randomized controlled clinical trials: what exactly are we measuring? *The Journal of investigative dermatology*. 2003;120(6):932-41.
3. Schmitt J, Williams H, Group HD. Harmonising Outcome Measures for Eczema (HOME). Report from the First International Consensus Meeting (HOME 1), 24 July 2010, Munich, Germany. *The British journal of dermatology*. 2010;163(6):1166-8.
4. Clarke M. Standardising outcomes for clinical trials and systematic reviews. *Trials*. 2007;8:39.
5. Schmitt J, Apfelbacher C, Spuls PI, Thomas KS, Simpson EL, Furue M, et al. The Harmonizing Outcome Measures for Eczema (HOME) roadmap: a methodological framework to develop core sets of outcome measurements in dermatology. *The Journal of investigative dermatology*. 2015;135(1):24-30.
6. Chalmers JR, Simpson E, Apfelbacher CJ, Thomas KS, von Kobyletzki L, Schmitt J, et al. Report from the fourth international consensus meeting to harmonise core outcome measures for atopic eczema/dermatitis clinical trials (HOME initiative). *Br J Dermatol*. 2016;In press.
7. Gerbens LAA, Chalmers JR, Rogers NK, Nankervis H, Spuls PI. Reporting of symptoms in randomised controlled trials of atopic eczema treatments: a systematic review. *Br J Dermatol*. 2016;In press.
8. Moher D, Liberati A, Tetzlaff J, Altman D, Altman D, Antes G, et al. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med*. 2009;6(7).
9. Terwee CB, Jansma EP, Riphagen, II, de Vet HC. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation*. 2009;18(8):1115-23.
10. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation*. 2010;19(4):539-49.
11. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of clinical epidemiology*. 2007;60(1):34-42.
12. Furlan AD, Pennick V, Bombardier C, van Tulder M, Editorial Board CBRG. 2009 updated method guidelines for systematic reviews in the Cochrane Back Review Group. *Spine*. 2009;34(18):1929-41.
13. Schmitt J, Spuls P, Boers M, Thomas K, Chalmers J, Roekevisch E, et al. Towards global consensus on outcome measures for atopic eczema research: results of the HOME II meeting. *Allergy*. 2012;67(9):1111-7.
14. Boers M, Brooks P, Strand CV, Tugwell P. The OMERACT filter for Outcome Measures in Rheumatology. *The Journal of rheumatology*. 1998;25(2):198-9.
15. Dauden E, Sanchez-Perez J, Prieto M, Roset M. [Validation of the Spanish Version of the Itch Severity Scale: the PSEDA study]. *Actas dermo-sifiliograficas*. 2011;102(7):527-36.
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(12):1513-9.
17. Schram ME, Spuls PI, Leeflang MM, Lindeboom R, Bos JD, Schmitt J. EASI, (objective) SCORAD and POEM for atopic eczema: responsiveness and minimal clinically important difference. *Allergy*. 2012;67(1):99-106.
18. Charman CR, Venn AJ, Ravenscroft JC, Williams HC. Translating Patient-Oriented Eczema Measure (POEM) scores into clinical practice by suggesting severity strata derived using anchor-

based methods. *The British journal of dermatology*. 2013;169(6):1326-32.

19. Coutanceau C, Stalder JF. Analysis of correlations between patient-oriented SCORAD (PO-SCORAD) and other assessment scores of atopic dermatitis severity and quality of life. *Dermatology*. 2014;229(3):248-55.

20. Vourc'h-Jourdain M, Barbarot S, Taieb A, Diepgen T, Ambonati M, Durosier V, et al. Patient-oriented SCORAD: a self-assessment score in atopic dermatitis. A preliminary feasibility study. *Dermatology*. 2009;218(3):246-51.

21. Stalder JF, Barbarot S, Wollenberg A, Holm EA, De Raeve L, Seidenari S, et al. Patient-Oriented SCORAD (PO-SCORAD): a new self-assessment scale in atopic dermatitis validated in Europe. *Allergy*. 2011;66(8):1114-21.

22. Housman TS, Patel MJ, Camacho F, Feldman SR, Fleischer AB, Jr., Balkrishnan R. Use of the Self-Administered Eczema Area and Severity Index by parent caregivers: results of a validation study. *The British journal of dermatology*. 2002;147(6):1192-8.

23. Mazzotti E, Paradisi M, Provini A, Tabolli S. Reproducibility of severity rating for atopic dermatitis. *Acta dermato-venereologica*. 2008;88(1):60-1.

24. van Velsen SG, Knol MJ, Haeck IM, Bruijnzeel-Koomen CA, Pasmans SG. The Self-administered Eczema Area and Severity Index in children with moderate to severe atopic dermatitis: better estimation of AD body surface area than severity. *Pediatric dermatology*. 2010;27(5):470-5.

25. Charman D, Varigos G, Horne DJ, Oberklaid F. The development of a practical and reliable assessment measure for atopic dermatitis (ADAM). *Journal of outcome measurement*. 1999;3(1):21-34.

26. Charman DP, Varigos GA. Grades of severity and the validation of an atopic dermatitis assessment measure (ADAM). *Journal of outcome measurement*. 1999;3(2):162-75.

27. Darsow U, Scharein E, Simon D, Walter G, Bromm B, Ring J. New aspects of itch pathophysiology: component analysis of atopic itch using the 'Eppendorf Itch Questionnaire'. *International archives of allergy and immunology*. 2001;124(1-3):326-31.

28. Zachariae R, Lei U, Haedersdal M, Zachariae C. Itch severity and quality of life in patients with pruritus: preliminary validity of a Danish adaptation of the itch severity scale. *Acta dermato-venereologica*. 2012;92(5):508-14.

29. Haest C, Casaer M, Daems A, De Vos B, Vermeersch E, Morren M, et al. Measurement of itching: Validation of the Leuven Itch Scale. *Burns*. 2011;37(6):939-50.

30. Lob-Corzilius T, Boer S, Scheewe S, Wilke K, Schon M, Schulte Im Walde J, et al. The 'Skin Detective Questionnaire': A survey tool for self-assessment of patients with atopic dermatitis. First results of its application. *Dermatol Psychosom*. 2004;5(3):141-6.

31. Wu D, Huang C, Mo X, Liu J, Cai J, Liu C, et al. Development and initial validation of a Traditional Chinese Medicine symptom-specific outcome measure: a Zheng-related atopic dermatitis symptom questionnaire (ZRADSQ). *Health and quality of life outcomes*. 2013;11:212.

32. 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. *Annals of allergy, asthma & immunology : official publication of the American College of Allergy, Asthma, & Immunology*. 2008;101(5):500-7.

33. Dawn A, Papoiu AD, Chan YH, Rapp SR, Rassette N, Yosipovitch G. Itch characteristics in atopic dermatitis: results of a web-based questionnaire. *The British journal of dermatology*. 2009;160(3):642-4.

34. Yoshida H, Aoki T, Furue M, Tagami H, Kaneko F, Ohtsuka F, et al. English version of the interim report published in 1998 by the members of the Advisory Committee on Atopic Dermatitis Severity Classification Criteria of the Japanese Dermatological Association. *The Journal of dermatology*. 2011;38(7):625-31.

35. Emerson RM, Charman CR, Williams HC. The Nottingham Eczema Severity Score: preliminary refinement of the Rajka and Langeland grading. *The British journal of dermatology*. 2000;142(2):288-97.

36. Hon KL, Kam WY, Lam MC, Leung TF, Ng PC. CDLQI, SCORAD and NESS: are they correlated? Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation. 2006;15(10):1551-8.
37. Hon KL, Ma KC, Wong E, Leung TF, Wong Y, Fok TF. Validation of a self-administered questionnaire in Chinese in the assessment of eczema severity. Pediatric dermatology. 2003;20(6):465-9.
38. European Task Force on Atopic Dermatitis. Severity scoring of atopic dermatitis: the SCORAD index. Consensus Report of the European Task Force on Atopic Dermatitis. Dermatology. 1993;186(1):23-31.
39. Schafer T, Dockery D, Kramer U, Behrendt H, Ring J. Experiences with the severity scoring of atopic dermatitis in a population of German pre-school children. The British journal of dermatology. 1997;137(4):558-62.
40. Pucci N, Novembre E, Cammarata MG, Bernardini R, Monaco MG, Calogero C, et al. Scoring atopic dermatitis in infants and young children: distinctive features of the SCORAD index. Allergy. 2005;60(1):113-6.
41. Hon KL, Leung TF, Wong Y, Fok TF. Lesson from performing SCORADs in children with atopic dermatitis: subjective symptoms do not correlate well with disease extent or intensity. International journal of dermatology. 2006;45(6):728-30.
42. Rullo VE, Segato A, Kirsh A, Sole D. Severity scoring of atopic dermatitis: a comparison of two scoring systems. Allergologia et immunopathologia. 2008;36(4):205-11.
43. Kido-Nakahara M, Katoh N, Saeki H, Mizutani H, Hagihara A, Takeuchi S, et al. Comparative cut-off value setting of pruritus intensity in visual analogue scale and verbal rating scale. Acta dermato-venereologica. 2015;95(3):345-6.
44. von Kobyletzki LB, Thomas KS, Schmitt J, Chalmers JR, Deckert S, Aoki V, et al. What is important to patients when assessing treatment response: an international cross-sectional survey. Acta dermato-venereologica. 2016;In press.
45. Melzack R. The McGill Pain Questionnaire: major properties and scoring methods. Pain. 1975;1(3):277-99.
46. Rajka G, Langeland T. Grading of the severity of atopic dermatitis. Acta dermato-venereologica Supplementum. 1989;144:13-4.

Tables

Table 1: Levels of evidence for the overall quality of a measurement property (12)

Level	Rating ¹	Criteria
Strong	+++ or ---	Consistent findings in multiple studies of good methodological quality or in one study of excellent methodological quality
Moderate	++ or --	Consistent findings in multiple studies of fair methodological quality or in one study of good methodological quality
Limited	+ or -	One study of fair methodological quality
Conflicting	±	Conflicting findings
Unknown	?	Only studies of poor methodological quality

¹ + = positive/good; ? = indeterminate; - = negative/poor; ± = conflicting rating. Findings on construct validity were combined if both convergent and/or discriminative and/or discriminant validity were assessed, following predefined criteria: if ratings were conflicting, the worst rating determined the overall rating; if either convergent, discriminative or discriminant validity had an indeterminate rating, the rating for the other component of construct validity determined the overall rating.

Table 2: Characteristics of included studies

Symptom instrument	Number of studies ¹	Study setting	Geographic location(s)	AE study population		
				Sample size per study (n)	Age (mean (SD) and/or range in years)	Severity of AE
ADAM	4 (25, 26)	Tertiary care	Australia	15 (25)	nr (nr), 0.3-15	nr
				171 (25)	4.5 (nr), 0.3-7.8	Mild 43%, moderate 48%, severe 7% based on global severity rating
				51 (25)	5.6 (4.5), 0.4-13.4	nr
				171 (26)	4.5 (nr), 0.3-7.8	nr
ADQ	1 (32)	Secondary/ Tertiary care	United States of America	68	4.6 (3.4), 0-12	Mild 3%, moderate 36%, severe 60% based on SCORAD
CoIQ	1 (33)	nr	United States of America	20, 15 healthy controls	nr	nr
EIQ	1 (27)	nr	Germany	108	nr (nr), 17-70	nr
ISS adult version	2 (15, 28)	Secondary care (15)	Spain	172, 35 controls	32.3 (13.4), nr	nr
		Tertiary care (28)	Denmark	20	31.4 (12.7), nr	SCORAD mean 40.4 (SD 13.1)
ISS paediatric version	1 (15)	Secondary care	Spain	151, 56 controls	8.1 (4.0), nr	nr
LIS	1 (29)	Secondary care	Belgium	63	33.4 (11.6), nr	nr
Method 4	1 (34)	Secondary care	Japan	259	21.8 (9.3), 0.5-58.2	Very mild 4.6%, mild 21.2%, moderate 47.5%, severe 20.5%, very severe 6.2% based on global evaluation
NESS	4 (35-37)	Secondary/ Tertiary care (35)	United Kingdom	nr	nr (nr), 1-5	nr
		Primary care (35)	United Kingdom	1761	nr (nr), 1-5	Mild 83%, moderate 15%, severe 2% based on dermatology assessment
		Tertiary care (37)	China	18, 52 caregivers	9.7 (nr), nr	nr
		Tertiary care (36)	China	80	11.7 (3.7), 5-20	SCORAD median 56.1
POEM	7 (16-19)	Primary and Secondary care (16)	United Kingdom	235	nr (nr), 1-58	nr
		Secondary care (16)	United Kingdom	200	nr (nr), 1-69 (median 9)	nr
		Secondary care (16)	United Kingdom	50	nr (nr), 1-62 (median 8.5)	nr
		Secondary care (16)	United Kingdom	40	nr (nr), 1-36 (median 4)	nr
		Tertiary care (17)	Germany, The Netherlands	80	34.9 (nr), nr	Moderate to severe
		Primary and Secondary	United Kingdom	1000	5.6 (nr), 0.08-65	POEM 13.6±7.1

		care (18)				
		Secondary and Tertiary care (19)	France, Poland, Italy, Spain, Belgium, Canada, Denmark, Czech Republic, Slovakia, Mexico, Romania, Greece	4222	11.9 (13.1), 0.1-97	Mild to moderate 68.7%, severe 31.3% based on SCORAD
PO-SCORAD	4 (19-21)	nr (20)	Europe	na	na	na
		Thermal spa (20)	France	33	22.7 (nr), 2-74	nr
		Secondary care (21)	Belgium, Denmark, Finland, France, Germany, Italy, The Netherlands, Sweden, Switzerland	471	20.9 (17.2) ² , 15.8 (16.6) ³	Mild to moderate 79.0%, severe 21.02% based on SCORAD
		Secondary and Tertiary care (19)	France, Poland, Italy, Spain, Belgium, Canada, Denmark, Czech Republic, Slovakia, Mexico, Romania, Greece	4222	11.9 (13.1), 0.1-97	Mild to moderate 68.7%, severe 31.3% based on SCORAD
Adapted SA-EASI	1 (19)	Secondary and Tertiary care (19)	France, Poland, Italy, Spain, Belgium, Canada, Denmark, Czech Republic, Slovakia, Mexico, Romania, Greece	4222	11.9 (13.1), 0.1-97	Mild to moderate 68.7%, severe 31.3% based on SCORAD
SA-EASI	3 (22-24)	Tertiary care (22)	United States of America	47	nr (nr), < 12	nr
		Secondary/ Tertiary care (23)	Italy	98	7.7 (5.1), 0.5-17	Moderate 61.2%, severe 38.8% based on SCORAD
		Tertiary care (24)	The Netherlands	60	10.5 (nr), 5-16	Moderate to severe
Subjective SCORAD	8 (27, 32, 36, 38-42)	Tertiary care (38)	Europe	na	na	na
		Primary care (39)	Germany	171	nr (nr), 5-6	SCORAD mean 23.6, 3.9 to 67.5
		nr (27)	Germany	108	nr (nr), 17-70	nr
		Secondary/ Tertiary care (40)	Italy	63	1.5 (nr), 0.2-4	Mild 19%, moderate 62%, severe 19% based on SCORAD
		Tertiary care (36)	China	80	11.7 (3.7), 5-20	SCORAD median 56.1
		Tertiary care (41)	China	182	9.6 (4.2), nr	SCORAD mean 30.1
		Secondary/ Tertiary care (32)	United States of America	68	4.6 (3.4), 0-12	Mild 3%, moderate 36%, severe 60% based on SCORAD
		Tertiary care (42)	Brazil	42	nr	nr
SDQ	3 (30)	nr	Germany	na	na	na
		Secondary care	Germany	22	11.7 (nr), 7-21	nr
		Primary/ Secondary care	Germany	324	nr (nr), 8-18	SCORAD > 20
VAS pruritus	1 (43)	Tertiary care	Japan	181	nr	nr

VRS pruritus	1 (43)	Tertiary care	Japan	181	nr	nr
ZRADSQ	2 (31)	Tertiary care	China	40	nr (nr), 7-25	nr
		Tertiary care	China	188	13.6 (nr), 5-25	Moderate to severe

Abbreviations: ADAM, Atopic Dermatitis Assessment Measure; ADQ, Atopic Dermatitis Quickscore; ColQ, Web-based Characteristics of itch questionnaire; EIQ, Eppendorf Itch Questionnaire; ISS, Itch Severity Scale; LIS, Leuven Itch Scale; na, not applicable; NESS, Nottingham Eczema Severity Score; nr, not reported; POEM, Patient-Oriented Eczema Measure; PO-SCORAD, Patient-Oriented SCORing Atopic Dermatitis index; SA-EASI, Self-administered Eczema Area and Severity Index; SCORAD, SCORing Atopic Dermatitis index; SD, standard deviation; SDQ, Skin Detective Questionnaire; VAS, Visual Analogue Scale; VRS, Verbal Rating Scale; ZRADSQ, Zheng-related atopic dermatitis symptom questionnaire.

¹ Different studies may be described in one paper

² Patients with severe AE

³ Patients with mild to moderate AE

Table 3: Characteristics of included instruments

Instrument	Kind of symptoms assessed	Other domains assessed	Target population	Assessor	Number of components (c) and/or items (i)	Method of rating	Scoring algorithm	Interpretation of total score	Available translations	Completion time (minute)
ADAM	Pruritus (severity)	Signs (site & morphology), global rating of severity	AE	Patient and physician	3 (c), 29 (i), 1 (i) for symptoms	2- and 4-point VRS	Sum score, range 0-46	0-2= trivial, 3-14= mild, 15-31= moderate, 32-46= severe	English	nr
ADQ	Skin and pruritus severity on separate body parts	na	AE	Parent/caregiver	2 (c), 14 (i)	6-point VRS	Sum score, range 0-70	nr	English	< 5
ColQ	Pruritus (frequency, location, duration, intensity, effect of scratching, associated symptoms like pain, sweating, heat sensation, cold sensation)	na	AE (based on EIQ)	Patient	32 (i)	5-point VRS for itch descriptors, frequency, duration, 10-point NRS for intensity, 11-point NRS for effect of scratching	nr	nr	English	15-20
EIQ	Pruritus (descriptors of itch, location, diurnal and reaction items)	na	AE (based on McGill pain questionnaire (45))	Patient	130 (i)	5-point NRS for itch descriptors, VAS for intensity	nr	nr	English	nr

ISS	Pruritus (frequency, itch description, affected body surface area, intensity, effect on sleep, effect on mood, effect on sexual desire/function (<i>omitted from paediatric version</i>))	na	Patients with pruritus (based on EIQ)	Patient or parent/caregiver	7 (c), 24 (i)	VRS	Scores of each component converted to standardised score (actual score divided by maximum score), ranging from 0.0 to 1.0. Standardised scores summed and multiplied by 3. Range 0-21.	0 (no itch) to 21 (very severe itching)	Danish, Spanish	nr
LIS	Pruritus (symptom occurrence (frequency, duration, severity, circumstances), distress, management, location, sensory perception, consequences)	na	Patients with itch	Patient	11 (c)	VRS, VAS, check box, drawing	nr	nr	Dutch, English	nr
Method 4	Severity of daytime pruritus, sleep disturbance	na	AE	Patient	2 (i)	5-point VRS	nr	nr	English, Japanese	nr
NESS	Clinical course, clinical intensity by sleep disturbance due to itching and scratching, extent	na	AE (based on Rajka and Langeland index (46))	Patient or parent	3 (i)	5-point VRS and NRS	Sum score, range 3-15	3-8= mild, 9-11= moderate, 12-15= severe	English, Chinese	1
POEM	Frequency of pruritus, sleep disturbance, bleeding, weeping or oozing, cracking, flaking, dryness or roughness	na	AE	Patient or parent/caregiver	7 (i)	5-point VRS	Sum score, range 0-28	0-2= (almost) clear, 3-7= mild, 8-16= moderate, 17-24= severe, 25-28= very severe	10 (nottingham.ac.uk)	1-2
PO-SCORAD	Affected surface area; severity of dryness, redness, swelling, crusting/oozing, scratching, thickening; severity of itching, sleep	na	AE (based on SCORAD index)	Patient or parent/caregiver	3 (c), 9 (i)	Component 1: shading and describing, physician estimates %; component 2: 4-point NRS; component 3:	nr	nr	19 (poscorad.com)	< 5 (48%), < 1 (96%)

	disturbance					VAS				
Adapted SA-EASI	Severity of redness, thickness, dryness, number of scratches, itchiness	na	AE (based on SA-EASI)	Patient or parent/caregiver	5 (i)	100mm VAS	nr	nr	English	nr
SA-EASI	Affected surface area; severity of redness, thickness, dryness, number of scratches, itchiness	na	AE (based on EASI)	Patient or parent/caregiver	2 (c), 6 (i)	Component 1: 4 areas of the body rated on 7-point NRS; component 2: 100mm VAS	Formulas: component 1 (each area score multiplied by multiplier assigned to corresponding body area) and component 2 (subcategorised into acute and chronic manifestations), adjusted for BSA differences in children > 7 yrs. vs. ≤ 7 yrs.	nr	English, Italian	nr
Subjective SCORAD (part C of SCORAD index)	Intensity of pruritus and sleep loss	na	AE	Patient	2 (i)	10cm VAS	Sum score (in case of SCORAD index: A/5+7*B/2+C)	Total SCORAD index: <15= mild, 15-40= moderate, > 40= severe	English	≤ 10 (total SCORAD inde
SDQ	Extent; severity of dryness, redness, knotty swellings or small blisters, weeping or scabbing, scratching, deep creases of the skin/leather skin; intensity of itch and sleeplessness	na	AE (based on SCORAD index)	Patient	3 (c), 9 (i)	Sketching BSA, 4-point VRS, 0-10 NRS	nr	nr	English	8.25
VAS pruritus	Pruritus	na	Patients with pruritus	Patient	1 (i)	VAS	0-10	0= no pruritus, >0-<4= mild, ≥4-<7= moderate, ≥7-<9= severe, ≥9= very severe	nr but conducted in Japan	nr

VRS pruritus	Pruritus	na	Patients with pruritus	Patient	1 (i)	5-point VRS	0-5	0= no pruritus, 1= mild, 2= moderate, 3= severe, 4= very severe	nr but conducted in Japan	nr
ZRADSQ (Zheng represents pattern differentiation in Traditional Chinese Medicine)	AE symptoms (itching aggravated at night, itching, dry skin, itching accompanied by pain, burning skin, insomnia), heat (thirst, mouth dryness, constipation, dark urine), mood (fidgeting, irascibility)	na	AE	Patient or parent/ caregiver if < 7 years	15 (i)	4-point VRS, 10cm VAS	Sum score	nr	Chinese, English	nr

Abbreviations: ADAM, Atopic Dermatitis Assessment Measure; ADQ, Atopic Dermatitis Quickscore; AE, Atopic eczema; BSA, Body surface area; CoIQ, Web-based Characteristics of itch questionnaire; EIQ, Eppendorf Itch Questionnaire; ISS, Itch Severity Scale; LIS, Leuven Itch Scale; na, not applicable; NESS, Nottingham Eczema Severity Score; nr, not reported; POEM, Patient-Oriented Eczema Measure; PO-SCORAD, Patient-Oriented SCOring Atopic Dermatitis index; SA-EASI, Self-administered Eczema Area and Severity Index; SCORAD, SCOring Atopic Dermatitis index; SDQ, Skin Detective Questionnaire; VAS, Visual Analogue Scale; VRS, Verbal Rating Scale; ZRADSQ, Zheng-related atopic dermatitis symptom questionnaire.

Table 4: Methodological quality of included studies according to the COSMIN checklist

	IRT	Box A	Box B	Box C	Box D	Box E	Box F	Box G	Box I	Box J
Instrument		Internal consistency	Reliability	Measurement error	Content validity	Structural validity	Hypotheses testing ¹	Cross-cultural validity	Responsiveness	Interpretability
ADAM substudy 1 (25)					poor					
substudy 2 (25)						good				
substudy 3 (25)			fair							
(26)	good	fair					fair			yes
ADQ (32)							fair		poor	
ColQ (33)		poor	fair		poor					
EIQ (27)						fair	fair			
ISS adult version (15)		fair	fair			fair	fair		fair	yes
(28)							fair	poor		
ISS paediatric version (15)		fair	fair			fair	fair		fair	yes
LIS (29)			fair		excellent	poor	fair			yes
Method 4 (34)							fair			
NESS substudy 1 (35)					fair					
substudy 2 (35)							fair			
(37)			poor	poor				poor		
(36)							fair			
POEM substudy 1 (16)					good					
substudy 2 (16)		fair					fair			
substudy 3 (16)			poor	fair						
substudy 4 (16)									fair	
(17)				poor					fair	yes
(18)										yes
(19)							fair		fair	yes
PO-SCORAD substudy 1 (20)					fair					
substudy 2 (20)							fair			
(21)							fair			
(19)							fair		fair	yes
Adapted SA-EASI (19)							fair		fair	yes
SA-EASI (22)					poor		fair			
(23)			fair				poor			
(24)							fair			
Subjective SCORAD (38)					poor					
(39)						excellent				
(27)							fair			
(40)							poor			
(36)							fair			
(41)							fair			
(32)							fair		poor	
(42)			fair							
SDQ substudy 1 (30)					poor					
substudy 2 (30)			poor							
substudy 3 (30)			fair							

VAS pruritus (43)							fair			
VRS pruritus (43)							fair			
ZRADSQ substudy 1 (31)					poor					
substudy 2 (31)		fair	poor			fair	poor		poor	yes

Abbreviations: ADAM, Atopic Dermatitis Assessment Measure; ADQ, Atopic Dermatitis Quickscore; ColQ, Web-based Characteristics of itch questionnaire; EIQ, Eppendorf Itch Questionnaire; IRT, Item Response Theory; ISS, Itch Severity Scale; LIS, Leuven Itch Scale; NESS, Nottingham Eczema Severity Score; POEM, Patient-Oriented Eczema Measure; PO-SCORAD, Patient-Oriented SCORing Atopic Dermatitis index; SA-EASI, Self-administered Eczema Area and Severity Index; SCORAD, SCORing Atopic Dermatitis index; SDQ, Skin Detective Questionnaire; VAS, Visual Analogue Scale; VRS, Verbal Rating Scale; ZRADSQ, Zheng-related atopic dermatitis symptom questionnaire.

¹ Subtypes of hypothesis testing include convergent validity, divergent validity, discriminant validity and discriminative validity.

Table 5: Quality assessment of measurement properties according to predefined criteria proposed by Terwee (11)

	Box A	Box B	Box C	Box D	Box E	Box F	Box G	Box I	Box J
Instrument	Internal consistency	Reliability	Measurement error	Content validity	Structural validity	Hypotheses testing	Cross-cultural validity	Responsiveness	Interpretability ¹
ADAM (25) substudy 1				?					
(25) substudy 2					-				
(25) substudy 3		-							
(26)	+					nr			+
ADQ (32)						+		?	
ColQ (33)	+	+		?					
EIQ (27)					+	-			
ISS adult version (15)	+	-			+	+		?	+
(28)						-	?		
ISS paediatric version (15)	+	+			+	+		?	+
LIS (29)		-		+	?	-			+
Method 4 (34)						?			
NESS (35) substudy 1				?					
(35) substudy 2						?			
(37)		+	?				?		
(36)						?			
POEM (16) substudy 1				+					
(16) substudy 2	+					+			
(16) substudy 3		?	?						
(16) substudy 4								?	
(17)			-					+	+
(18)									+
(19)						+		+	+
PO-SCORAD (20) substudy 1				+					
substudy 2 (20)						-			
(21)						+			
(19)						+		+	+
Adapted SA-EASI (19)						+		+	+
SA-EASI (22)				?		+			
(23)		+				+			
(24)						±			
Subjective SCORAD (38)				?					
(39)					+				
(27)						-			

(40)						-			
(36)						-			
(41)						-			
(32)						+		?	
(42)		?							
SDQ (30) substudy 1				?					
(30) substudy 2		?							
(30) substudy 3		-							
VAS pruritus (43)						?			
VRS pruritus (43)						?			
ZRADSQ (31) substudy 1				?					
(31) substudy 2	+	+			-	?		?	+

Abbreviations: ADAM, Atopic Dermatitis Assessment Measure; ADQ, Atopic Dermatitis Quickscore; ColQ, Web-based Characteristics of itch questionnaire; EIQ, Eppendorf Itch Questionnaire; ISS, Itch Severity Scale; LIS, Leuven Itch Scale; NESS, Nottingham Eczema Severity Score; nr, not reported; POEM, Patient-Oriented Eczema Measure; PO-SCORAD, Patient-Oriented SCORing Atopic Dermatitis index; SA-EASI, Self-administered Eczema Area and Severity Index; SCORAD, SCORing Atopic Dermatitis index; SDQ, Skin Detective Questionnaire; VAS, Visual Analogue Scale; VRS, Verbal Rating Scale; ZRADSQ, Zheng-related atopic dermatitis symptom questionnaire.

+ = positive rating; ? = indeterminate rating; - = negative rating

¹ Quality rating for interpretability not according to COSMIN / Terwee *et al.* since no formal quality criteria for interpretability exist.

Table 6: Best evidence synthesis and recommendations

	Box A	Box B	Box C	Box D	Box E	Box F	Box G	Box I	Recommendation
Instrument	Internal consistency	Reliability	Measurement error	Content validity	Structural validity	Hypotheses testing	Cross-cultural validity	Responsiveness	
ADAM	+	-		?	--	n/i			C
ADQ						+		?	D
ColQ	?	+		?					D
EIQ					+	-			C
ISS adult version	+	-			+	±	?	? limited	C
ISS paediatric version	+	+			+	+		? limited	B
LIS		-		+++	?	-			C
Method 4						? limited			D
NESS		?	?	? limited		? limited	?		D
POEM	+	?	? limited	++		++		++	B
PO-SCORAD				+		±		+	B
Adapted SA-EASI						+		+	B
SA-EASI		+		?		+			B
Subjective SCORAD		? limited		?	+++	±		?	D
SDQ		-		?					C
VAS pruritus						? limited			D

VRS pruritus						? limited			D
ZRADSQ	+	?		?	-	?		?	C

Abbreviations: ADAM, Atopic Dermatitis Assessment Measure; ADQ, Atopic Dermatitis Quickscore; ColQ, Web-based Characteristics of itch questionnaire; EIQ, Eppendorf Itch Questionnaire; ISS, Itch Severity Scale; LIS, Leuven Itch Scale; NESS, Nottingham Eczema Severity Score; POEM, Patient-Oriented Eczema Measure; PO-SCORAD, Patient-Oriented SCORing Atopic Dermatitis index; SA-EASI, Self-administered Eczema Area and Severity Index; SCORAD, SCORing Atopic Dermatitis index; SDQ, Skin Detective Questionnaire; VAS, Visual Analogue Scale; VRS, Verbal Rating Scale; ZRADSQ, Zheng-related atopic dermatitis symptom questionnaire.

+++, ++, + = positive rating indicating adequate measurement property; ?, ? limited = indeterminate rating indicating indeterminate measurement property; ---, --, - = negative rating indicating inadequate measurement property; ± = conflicting findings; n/i = not interpretable

Supplementary Material

Detailed results category C and D instruments

Atopic Dermatitis Assessment Measure (ADAM)

Four studies evaluated the measurement properties of the ADAM (Charman *et al.*, 1999; Charman and Varigos, 1999). There was limited evidence for good internal consistency (Charman and Varigos, 1999) and for poor inter-rater reliability (Charman *et al.*, 1999). There was unknown evidence for content validity since the methodological quality of the content validity study was poor (Charman *et al.*, 1999). There was moderate evidence for poor structural validity (Charman *et al.*, 1999). The results for hypotheses testing were not interpretable since authors mentioned the assessment but results were not reported (Charman and Varigos, 1999). Interpretation of scores was assessed and yielded four grades of severity (i.e. trivial, mild, moderate, severe) (Charman and Varigos, 1999).

Atopic Dermatitis Quickscore (ADQ)

Only one study was found on the measurement properties of the ADQ, an instrument developed for caregivers of patients with AE (Carel *et al.*, 2008). There was limited evidence for good construct validity (i.e. convergent validity). The evidence for responsiveness is unknown as the methodological quality of the study was poor.

Web-based Characteristics of Itch Questionnaire (CoIQ)

One study was performed (Dawn *et al.*, 2009). Limited evidence was found for good test-retest reliability. There was unknown evidence for internal consistency and content validity since the methodological quality of the study was poor.

Eppendorf Itch Questionnaire (EIQ)

One study on the measurement properties of the EIQ was conducted (Darsow *et al.*, 2001). There was limited evidence for good structural validity and poor construct validity.

Adult Itch Severity Scale (ISS)

Two studies evaluated the measurement properties of the ISS (Dauden *et al.*, 2011; Zachariae *et al.*, 2012). Limited evidence was found for good internal consistency and for poor test-retest reliability for the Spanish ISS (Dauden *et al.*, 2011). Good structural validity was supported by limited evidence for the Spanish ISS (Dauden *et al.*, 2011). Conflicting findings were identified for construct validity for the Spanish and Danish ISS (Dauden *et al.*, 2011; Zachariae *et al.*, 2012). Evidence for cross-cultural validity of the Danish ISS is unknown since the methodological quality of the study was poor (Zachariae *et al.*, 2012). There was limited evidence for indeterminate responsiveness for the Spanish ISS (Dauden *et al.*, 2011). Interpretability yielded a minimal important difference (MID) of 4 points for the Spanish ISS (Dauden *et al.*, 2011).

Leuven Itch Scale (LIS)

One study assessed the measurement properties of the LIS (Haest *et al.*, 2011). Limited evidence was identified for poor test-retest reliability and for poor convergent and divergent validity. Strong

evidence was found for good content validity. There was unknown evidence for structural validity due to poor methodological quality of the study. No floor and ceiling effects were discovered.

Method 4

One study provided limited evidence for indeterminate construct validity (Yoshida *et al.*, 2011).

Nottingham Eczema Severity Score (NESS)

Four studies were found on the measurement properties of the NESS (Emerson *et al.*, 2000; Hon *et al.*, 2006a; Hon *et al.*, 2003). There was unknown evidence for inter-rater reliability, measurement error and cross-cultural validity as the methodological study quality was poor for the Chinese NESS (Hon *et al.*, 2003). Limited evidence was found for indeterminate content validity of the English NESS (Emerson *et al.*, 2000) and construct validity of both the Chinese and English NESS (Emerson *et al.*, 2000; Hon *et al.*, 2006a).

Subjective SCORing Atopic Dermatitis (SCORAD)

The measurement properties of the subjective part of the SCORAD index (part C) were addressed separately from the total SCORAD index in eight studies (Carel *et al.*, 2008; Darsow *et al.*, 2001; European Task Force on Atopic Dermatitis, 1993; Hon *et al.*, 2006a; Hon *et al.*, 2006b; Pucci *et al.*, 2005; Rullo *et al.*, 2008; Schafer *et al.*, 1997). Limited evidence was found for indeterminate inter-rater reliability (Rullo *et al.*, 2008). There was unknown evidence for content validity (European Task Force on Atopic Dermatitis, 1993) and responsiveness (Carel *et al.*, 2008) since the study quality was poor. Strong evidence was found for good structural validity of the total SCORAD index (Schafer *et al.*, 1997). Principal component analysis demonstrated, however, that the two items of the subjective SCORAD belong to different factors, indicating that part C is not unidimensional (i.e. the items do not measure the same construct). Therefore, calculating a sum score of the subjective part is not correct and, moreover, not possible to interpret. Five studies found evidence for construct validity, four with conflicting findings (Carel *et al.*, 2008; Darsow *et al.*, 2001; Hon *et al.*, 2006a; Hon *et al.*, 2006b) and unknown findings from one study of poor methodological quality (Pucci *et al.*, 2005).

Skin Detective Questionnaire (SDQ)

Three studies evaluated the measurement properties of the SDQ (Lob-Corzilius *et al.*, 2004). One study found limited evidence for poor inter-rater reliability and one study showed unknown evidence due to poor methodological quality. Unknown evidence for content validity was found since the methodological quality was poor.

Visual Analogue Scale (VAS) pruritus

One study provided limited evidence for indeterminate construct validity (Kido-Nakahara *et al.*, 2015).

Verbal Rating Scale (VRS) pruritus

Limited evidence in one study was found for indeterminate construct validity (Kido-Nakahara *et al.*, 2015).

Zheng-Related Atopic Dermatitis Symptom Questionnaire (ZRADSQ)

Two studies reported on the measurement properties of the ZRADSQ (Wu *et al.*, 2013). Limited evidence was found for good internal consistency. Unknown evidence was shown for test-retest reliability, content validity, convergent and discriminative validity and responsiveness since the methodological study quality was poor. There was limited evidence for poor structural validity. Interpretability demonstrated no floor and ceiling effects.

Figure S1: Flow diagram of identification and selection process (according to the PRISMA statement)

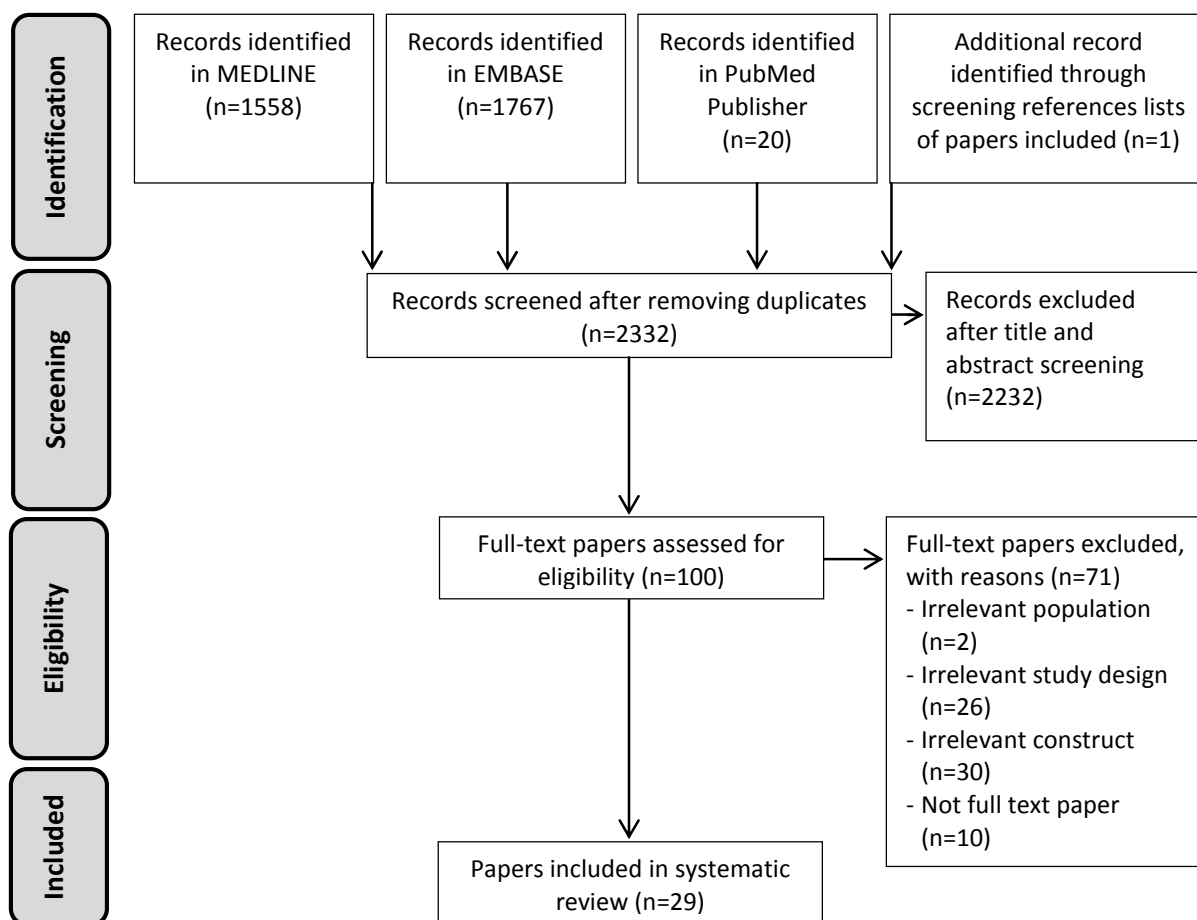


Table S1: Search strategy

Database(s): **Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations and Ovid MEDLINE(R) 1946 to Present**
 Search Strategy: **2015-08-03**

#	Searches	Results
1	Dermatitis, Atopic/	15134
2	exp Eczema/	9226
3	eczem*.tw,ot,kf.	15653
4	(atopic adj5 (dermatit* or neurodermatit*)).tw,ot,kf.	14853
5	(disseminated adj2 neurodermatit*).tw,ot,kf.	35
6	or/1-5	34340
7	(animals/ or (canine or dog or dogs or cat or cats).ti.) not humans/	4011041
8	editorial.pt.	391473
9	6 not (7 or 8)	32264
10	((exp Health Status Indicators/ or exp "Outcome Assessment (Health Care)"/ or Comparative Study/ or (instrumentation or methods).fs.) not case reports.pt.) or validation studies.pt. or observer variation/ or discriminant analysis/ or Psychometrics/ or "Reproducibility of Results"/ or factor analysis, statistical/ or evaluation studies/ or (audit or audits or psychometr* or clin?metr* or ((outcome* or clinical or observer* or utility or satisfaction or QoL or quality of life or score or scores or method or methods or physicians or modelling or objective) adj3 assessm*) or clinical asses* or outcome measure* or observer variation* or reproducib* or reliab* or unreliab* or valid* or coefficient or homogeneity or homogeneous or ((internal or external) adj3 (consistency or inconsistency)) or cronbach* or (item and (correlation* or selection* or reduction*)) or ((item or items) adj3 (discriminant* or convergent* or divergent*)) or agreement or precision or imprecision or (precise adj values) or (test and retest) or accuracy test* or stability or interrater or intrarater or intertester or intratester or interobserver or intraobserver or intertechnician or intratechnician or interexaminer or intraexaminer or interassay or intraassay or interindividual or intraindividual or interparticipant or intraparticipant or ((inter or intra) adj (rater or tester or observer or technician or examiner or assay or individual or participant)) or kappa or kappa's or kappas or repeatab* or ((replicab* or repeated) and (measure or measures or findings or result or results or test or tests)) or generaliza* or generalisa* or concordance or (intraclass and correlation*) or discriminative or (known adj group) or (factor adj (analy* or structure*)) or dimension* or interscale or inter-scale or interscales or inter-scales or subscale* or sub-scale* or ((multitrait* or multi-trait*) and (scaling or scale*)) or error or errors or ((individual or interval or rate) adj variability) or (variability adj5 (analy* or values)) or (uncertainty and (measurement or measuring)) or sensitiv* or responsive* or ((limit or limits) and detection) or ((minim* or lowest) adj2 detectable adj2 (concentration* or dose* or level* or amount*)) or interpretab* or (small* and (real or detectable) and (change or difference)) or meaningful change* or ((minimal* or minimum) adj2 (meaningful or important or detectable or real or identifiable or relevant) adj3 (change* or difference* or improvement*)) or ((minimal or minimally or clinical or clinically) and (important or significant) and (change* or difference* or improvement*)) or (MDC adj2 value*) or MCID or MCIDs or MICD or MICDs or MCII or MCIC or MCICs or ((ceiling or floor) adj2 effect*) or item response* or IRT or rasch or ((differential or fit) adj2 item*) or DIF or computer adaptive test* or item bank* or cross-cultural equivalen*).tw,ot,kf. [COSMIN-filter]	7249424
11	Feasibility Studies/	47828
12	(feasibilit* or practicab* or practicalit* or practibil* or intuitiv* or accept?bility or workab* or viability or expedien* or usefulness or (complet* adj2 time) or quality criter*).tw,ot,kf.	378101
13	((easy or ease or fast or simple or practical or feasible) adj6 ("use" or apply or perform* or method* or measure or measurement or outcome or model or instrument* or tool or index or indices or score* or scoring or scale* or subscale* or assess* or evaluat* or fill in or value)).tw,ot,kf.	262320
14	or/10-13 [COSMIN & FEASIBILITY FILTER]	7501503
15	*Severity of Illness Index/	14295

16	Patient Participation/	19154
17	Self-Assessment/ or Self-Examination/	11636
18	POEM.tw,ot,kf.	605
19	((self or patient*) adj (orient* or assess* or important or relevant or centered or defined or scale* or subscal* or scor* or subscor* or rating* or rate* or index or indices)).tw,ot,kf.	53627
20	((self or patient*) adj reported adj (outcome* or benefit* or symptom* or scale* or subscal* or scor* or subscor* or rating* or rate or index or indices or assess* or item* or questionn* or sensation* or perceiv* or percept* or feel* or intensit* or severit*)).tw,ot,kf.	10767
21	self-reporting.tw,ot,kf.	2557
22	(SCORAD or mSCORAD).tw,ot,kf.	623
23	(SCORing adj2 (Atopic Dermati* or AD)).tw,ot,kf.	381
24	(SCORing adj ind*).tw,ot,kf.	337
25	(ADSI or ADASI).tw,ot,kf.	17
26	(EASI or mEASI or mLEASI).tw,ot,kf.	285
27	((severity or disease severit* or sensitivity) adj2 (scor* or subscor* or scale or scales or subscale* or index* or indices)).tw,ot,kf.	32800
28	PAIS.tw,ot,kf.	947
29	(individual adj2 (sign or signs or symptom*)).tw,ot,kf.	2329
30	(SASSAD or mSASSAD).tw,ot,kf.	36
31	Six area six sign.tw,ot,kf.	36
32	Pain Measurement/	65346
33	(VAS or PVAS or VASP or (visu* adj3 analo* adj3 (scal* or scor*))).tw,kw.	51648
34	perception/ or exp pain perception/	24748
35	Pain Threshold/	10095
36	pruritis/di	726
37	((prurit* or rash* or itch* or scratching or irritat* or burning or prickling or pain or stinging or smarting or sore* or dry skin or skin symptom* or dryness or sleep loss* or sleep*) adj3 (scale* or scor* or rating* or rate or subscor* or analog* or test* or measur* or assess* or item* or questionn* or report* or sensation* or perceiv* or percept* or feel* or visual* or subjectiv* or intensit* or severit*)).tw,ot,kf.	121699
38	((symptom* or sign or signs or disease activity) adj3 (scale* or subscal* or scor* or subscor* or rating* or rate or index or indices or assess* or item* or questionn* or sensation* or perceiv* or percept* or feel* or intensit* or severit*)).tw,ot,kf.	91701
39	subjectiv*.ti,ot,kf.	11966
40	(subjectiv* adj5 (symptom* or complaint* or scale* or subscal* or scor* or subscor* or rating* or rate* or index or indices or item* or tool* or instrument* or perception* or assess* or measur* or test* or report* or severity)).tw,ot,kf.	46778
41	or/15-40 [outcome measurement tools]	447921
42	41 and 9 and 14 [I = AD symptom measurement tools + filters]	1540
43	(SCORAD or POEM or PAIS or SASSAD or mSASSAD or six area sign or EASI or mLEASI or ADSI or ADASI or ((scoring or score* or subscor* or quickscor* or severity ind* or subjective or patient-orient* or ((VAS or visu* analo*) and (sleep* or itch or prurit*))) and (dermati* or eczem* or AD))).ti. [AD + symptom measurement tools TITLE SEARCH]	469
44	43 and 9 [II = AD-symptom measurement tools in TITLE]	109
45	42 or 44 [I & II]	1558

Table S2: In- and exclusion criteria of study selection

	Inclusion criteria	Exclusion criteria
Population	AE (synonyms: atopic eczema, eczema, atopic dermatitis, neurodermatitis)	Other skin diseases including other forms of eczema (e.g. chronic hand eczema), <50% of population has AE unless subgroup analyses are present
Study design	Development or validation study	Pure linguistic validation studies
Outcome domain	Symptoms	Exclusively clinical signs, quality of life, long-term control of flares, biomarker, and physiology of the skin
Type of instrument	Patient-reported or by proxy (i.e. by parent, caregiver or interview)	All others
Type of publication	Papers with available full text	Abstracts

Table S3: Definitions of domains, measurement properties and aspects of measurement properties

Domain	Measurement property	Aspect of measurement property	Definition
Reliability			The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions
	Internal consistency		The degree of interrelatedness among the items
	Reliability		The proportion of the total variance in the measurements which is due to 'true' ¹ difference between patients
	Measurement error		The systematic and random error of a patient's score that is not attributed to true change of the construct to be measured
Validity			The degree to which a measurement instrument truly measures the construct(s) it purports to measure
	Content validity		The degree to which the content of an instrument is an adequate reflection of the construct to be measured
		Face validity	The degree to which a measurement instrument, indeed, looks as though it is an adequate reflection of the construct to be measured
	Construct validity		The degree to which the scores of a measurement instrument are consistent with hypotheses, e.g. with regard to internal relationships, relationships to scores of other instruments, or differences between relevant groups) based on the assumption that the instrument validly measures the construct to be measured
		Structural validity	The degree to which the scores of an instrument are an adequate reflection of the dimensionality of the construct to be measured
		Hypothesis testing	Idem construct validity
		Cross-cultural validity	The degree to which the performance of the items on a translated or culturally adapted instrument are an adequate reflection of the performance of items in the original version of the instrument
Responsiveness			The ability of a measurement instrument to detect change over time in the construct to be measured
	Responsiveness		Idem responsiveness
Interpretability ²			The degree to which one can assign qualitative meaning – that is, clinical or commonly understood connotations – to an instrument's quantitative scores or changes in scores

¹ The word 'true' must be seen in the context of the CCT (classical test theory), which states that any observation is composed of two components – a true score and error associated with the observation. 'True' is the average score that would be obtained if the scale were given an infinite number of times. It refers only to the consistency of the score and not to its accuracy.

² Interpretability is not considered a measurement property but an important characteristic of a measurement instrument.

Table S4: Quality criteria for measurement properties (Terwee *et al.*, 2007)

Property	Rating ¹	Quality Criteria
Reliability		
Internal consistency	+	Cronbach's alpha(s) ≥ 0.70
	?	Cronbach's alpha not determined or dimensionality unknown
	-	Cronbach's alpha(s) < 0.70
Reliability	+	ICC / weighted Kappa ≥ 0.70 OR Pearson's r ≥ 0.80
	?	Neither ICC / weighted Kappa, nor Pearson's r determined
	-	ICC / weighted Kappa < 0.70 OR Pearson's r < 0.80
Measurement error	+	MIC > SDC OR MIC outside the LOA
	?	MIC not defined
	-	MIC $\leq$ SDC OR MIC equals or inside LOA
Validity		
Content validity	+	All items are considered to be relevant for the construct to be measured, for the target population, and for the purpose of the measurement AND the questionnaire is considered to be comprehensive
	?	Not enough information available
	-	Not all items are considered to be relevant for the construct to be measured, for the target population, and for the purpose of the measurement OR the questionnaire is considered not to be comprehensive
Construct validity - Structural validity	+	Factors should explain at least 50% of the variance OR CFI > 0.95 AND RMSEA < 0.06
	?	Explained variance not mentioned OR important statistics not reported
	-	Factors explain < 50% of the variance OR CFI ≤ 0.95 OR RMSEA > 0.06
- Hypothesis testing	+	Correlations with instruments measuring the same construct ≥ 0.50 OR at least 75% of the results are in accordance with the hypotheses AND correlations with related constructs are higher than with unrelated constructs
	?	Solely correlations determined with unrelated constructs
	-	Correlations with instruments measuring the same construct < 0.50 OR < 75% of the results are in accordance with the hypotheses OR correlations with related constructs are lower than with unrelated constructs
- Cross-cultural validity	+	No differences in factor structure OR no important DIF between language versions
	?	Multiple group factor analysis not applied AND DIF not assessed
	-	Differences in factor structure OR important DIF between language versions
Responsiveness		
Responsiveness	+	Correlation with changes on instruments measuring the same construct ≥ 0.50 OR at least 75% of the results are in accordance with the hypotheses OR AUC ≥ 0.70 AND correlations with changes in related constructs are higher than with unrelated constructs
	?	Solely correlations determined with unrelated constructs
	-	Correlations with changes on instruments measuring the same construct < 0.50 OR < 75% of the results are in accordance with the hypotheses OR AUC < 0.70 OR correlations with changes in related constructs are lower than with unrelated constructs

Abbreviations: AUC, Area Under the Curve; CFI, Comparative Fit Index; DIF, Differential Item Functioning; ICC, Intraclass Correlation Coefficient; LoA, Limits of Agreement; MIC, Minimal Important Change; RMSEA, Root Means Square of Error of Approximation; SDC, Smallest Detectable Change.

¹ + = positive rating; ? = indeterminate rating; - = negative rating

Table S5: Named symptom instruments excluded with reasons

Instrument	Reason of exclusion (1,2,3)¹	Symptoms	Other domains
ADASI (Bahmer <i>et al.</i> , 1991)	1	Pruritus	Signs
ADSI (Van Leent <i>et al.</i> , 1998)	1	Pruritus	Signs
mEASI (Boguniewicz <i>et al.</i> , 1998)	3	Pruritus	Signs
PAIS (Brenninkmeijer <i>et al.</i> , 2010)	1	Pruritus	Signs
RESS (Fujitani <i>et al.</i> , 2014)	2	Overall severity, pruritus, sleep disturbance	-
SSRI (Xu <i>et al.</i> , 2008)	1	Pruritus	Signs
SSS (Costa's) (Costa <i>et al.</i> , 1989)	1	Pruritus, loss of sleep	Signs
SSS (Wu <i>et al.</i> , 2013b)	1	Pruritus	Signs
W-AZS (Silny <i>et al.</i> , 2005)	1	Pruritus, sleep disturbance	Signs

Abbreviations: ADASI, Atopic Dermatitis Area and Severity Index; ADSI, Atopic Dermatitis Severity Index; GCS, Global clinical score; mEASI, modified Eczema Area and Severity Index; PAIS, Physician Assessment of Individual Signs; RESS, Rapid Eczema Severity Score; SSRI, Symptom Score Reducing Index; SSS (Costa's), Simple Scoring System; SSS, Severity Scale Score; W-AZS, Polish for Index for Atopic Dermatitis.

¹ 1 = No validation of symptom measurement properties; 2 = Only abstract available; 3 = Content of instrument not clearly described and different versions available

Table S6: Pros and cons of category B instruments

Instrument	Pros	Cons
Paediatric ISS	-4 measurement properties with positive rating -Interpretability investigated	-1 measurement property with indeterminate evidence, 3 not investigated -Itch is the only symptom included -Feasibility unknown
POEM	-4 measurement properties with positive rating -Interpretability investigated -7 symptoms included -Feasibility (1-2 minutes to complete)	-2 measurement properties with indeterminate rating, 2 not investigated
PO-SCORAD	-2 measurement properties with positive rating -Interpretability investigated -8 symptoms included beside affected surface area -Feasibility (<10 min (48%), < 5 min (96%) to complete)	-1 measurement property with conflicting evidence, 5 not investigated
SA-EASI	-2 measurement properties with positive rating -5 symptoms included beside affected surface area	-1 measurement property with indeterminate rating, 5 not investigated -Feasibility unknown
Adapted SA-EASI	-2 measurement properties with positive rating -Interpretability investigated -5 symptoms included	-6 measurement properties not investigated -Feasibility unknown

Abbreviations: ISS, Itch Severity Scale; POEM, Patient-Oriented Eczema Measure; PO-SCORAD, Patient-Oriented SCORing Atopic Dermatitis index; SA-EASI, Self-administered Eczema Area and Severity Index.