Dermatol Ther (Heidelb) (2023) 13:2549–2571
https://doi.org/10.1007/s13555-023-01038-3

REVIEW

Assessing Response in Atopic Dermatitis: A Systematic Review of the Psychometric Performance of Measures Used in HTAs and Clinical Trials

Hannah Penton · Sayeli Jayade · Santhani Selveindran · Marieke Heisen · Christophe Piketty · Liliana Ulianov · Zarif K. Jabbar-Lopez · Jonathan I. Silverberg · Jorge Puelles

Received: August 1, 2023 / Accepted: September 11, 2023 / Published online: September 25, 2023

ABSTRACT

Introduction: Assessing treatment response is key to determining treatment value in atopic dermatitis (AD). Currently, response is assessed using various clinician- or patient-reported measures and response criteria. This variation creates a mismatch of evidence across trials, hindering the ability of clinicians, regulators, and payers to compare the efficacy of treatments. This review identifies which measures and criteria are used to determine response in clinical trials and health technology assessments (HTAs). Moreover, it systematically reviews the psychometric performance of those measures and criteria to understand which perform best in capturing patient-relevant symptoms and treatment benefits.

Methods: A scoping review of clinical trials and HTAs in AD identified the following measures for inclusion: the Eczema Area and Severity Index (EASI), the Investigator's Global Assessment (IGA), the Dermatology Life Quality Index (DLQI) and the Peak Pruritus Numerical Rating Scale (PP-NRS). A systematic search was performed in MEDLINE and Embase to identify studies testing the psychometric performance of these measures in adults or adolescents with AD.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13555-023-01038-3.

H. Penton (✉) · M. Heisen
OPEN Health, Rotterdam, The Netherlands
e-mail: HannahPenton@openhealthgroup.com

M. Heisen
e-mail: MariekeHeisen@openhealthgroup.com

S. Jayade
OPEN Health, Parsippany, NJ, USA
e-mail: SayeliJayade@openhealthgroup.com

S. Selveindran
OPEN Health, London, UK
e-mail: SanSelveindran@openhealthgroup.com

C. Piketty · L. Ulianov · Z. K. Jabbar-Lopez · J. Puelles
Galderma, La Tour-de-Peilz, Switzerland

C. Piketty
e-mail: Christophe.Piketty@galderma.com

L. Ulianov
e-mail: Liliana.Ulianov@galderma.com

Z. K. Jabbar-Lopez
e-mail: Zarif.Jabbar-Lopez@galderma.com

J. Puelles
e-mail: Jorge.Puelles@galderma.com

J. I. Silverberg
Department of Dermatology, The George Washington University School of Medicine and Health Sciences, Washington, DC, USA
e-mail: Jonathanisilverberg@gmail.com

Results: A lack of consistency in the assessment of response was observed across clinical trials and HTAs. Important gaps in psychometric evidence were identified. No content validations of the EASI and IGA in AD were found, while some quantitative studies suggested that these measures fail to capture itch, a core symptom. The PP-NRS and DLQI performed well. No studies compared the performance of different response criteria.

Conclusion: Content validation of the PP-NRS confirmed the importance of itch as a core symptom and treatment priority in AD; however, itch is not well covered in the EASI or IGA. Including the PP-NRS in clinical trials and HTAs will better capture patient-relevant benefit and response. Although various response criteria were used, no studies compared the performance of different criteria to inform which were most appropriate to compare treatments in clinical trials and HTAs.

PLAIN LANGUAGE SUMMARY

The assessment of treatment response is important in determining treatment value in atopic dermatitis (AD). This study aimed to identify which outcome measures and criteria are used to determine treatment response in clinical trials and health technology assessments (HTAs). The psychometric performance of identified outcome measures and criteria was then systematically reviewed to understand which perform best in capturing patient-relevant symptoms and treatment benefits in AD. The review identified and included the Eczema Area and Severity Index (EASI), Investigator's Global Assessment (IGA), Dermatology Life Quality Index (DLQI) and Peak Pruritus Numerical Rating Scale (PP-NRS) as response measures. Lack of consistency in how response is assessed across clinical trials and HTAs makes it difficult for clinicians and payers to compare the efficacies and cost-effectivenesses of different treatments and to make optimal treatment decisions. The review found that content validity (the extent to which a measure covers those symptoms and treatment benefits which are important to patients) was not assessed for EASI and IGA. EASI and IGA are often used to assess response in clinical trials and HTAs, but they miss key elements of the patient-relevant disease impact and treatment benefit, including itch. Treatments leading to improvements in missed symptoms (e.g. itch) will be undervalued using EASI and IGA, decreasing the chances of regulatory approval and reimbursement. Moreover, response criteria used in clinical trials and HTAs are sometimes adopted in prescriber settings. Here, if response assessment does not capture patient-relevant benefit, patients' access to tailored treatment may be restricted due to the perceived non-response.

Keywords: Atopic dermatitis; EASI; IGA; DLQI; PP-NRS; Health technology assessment; Response

Key Summary Points

A lack of consistency was observed in the assessment of treatment response in patients with atopic dermatitis (AD) both across clinical trials and between trials and health technology assessments (HTAs).

No content validations of the Eczema Area and Severity Index (EASI) and the Investigator's Global Assessment (IGA) were found, and mixed results were observed between these measures and measures of itch, which was identified as a core patient-relevant symptom.

Including the Peak Pruritus Numerical Rating Scale (PP-NRS) as a measure of itch in clinical trials and HTA will better capture patient-relevant benefit and response.

Studies comparing the psychometric performance of different response criteria are needed to inform which are appropriate to use to compare different treatments in AD.

INTRODUCTION

Atopic dermatitis (AD) is a chronic inflammatory disease characterized by inflamed eczematous skin, dryness, pruritus (itch), skin pain and excoriations. Globally, the estimated prevalence of AD is 1–3% in adults, with a two- to threefold increase in incidence in industrialized countries over the past decades [1]. Itch is a core symptom of AD and has a substantial impact on quality of life by causing self-consciousness, bleeding, problems with concentration and sleep disturbance [2]. Reducing itch is the most important treatment goal in patients with AD [3].

Due to advances in our understanding of AD and many unmet therapeutic needs, new therapies have been investigated, including interleukin inhibitors such as dupilumab, lebrikizumab, tralokinumab, and nemolizumab, as well as janus kinase inhibitors such as upadacitinib, baricitinib and abrocitinib [4]. The addition of new treatments to the rapidly changing AD treatment landscape means that regulators, health technology assessment (HTA) bodies and clinicians need to understand the comparative efficacies of new treatments compared to existing options. Defining and assessing treatment response is key to determining a treatment's comparative efficacy and value. Currently, in AD, there is no consensus on a standard outcome measure that should be used to assess response [5]. Response is assessed in trials using a variety of different clinician- or patient-reported measures and a range of different response criteria. This creates a mismatch of evidence across trials, hindering the ability of clinicians, regulators, and payers to directly compare the efficacies of different treatments.

It is vital that measures used to determine response are psychometrically valid in the population in which they are being used. Measures should be valid, reliable and responsive in the target population; they should be able to detect meaningful change as well as clinically relevant differences in change across treatment groups; and they should comprehensively capture the symptoms that are important to patients [6, 7]. To assess whether this is the case for response measures and response criteria used in AD, this literature review followed a two-step approach. First, a scoping review was conducted to identify which measures and criteria are being used to determine treatment response in patients with AD in clinical trials and HTAs. Second, the authors systematically reviewed the evidence on the psychometric performance of those measures and criteria identified in step one as being used to determine response in clinical trials and HTAs in AD. Through these steps, the literature review aimed to understand the extent to which the response measures and criteria being used in clinical trials and HTAs in AD comprehensively capture the symptoms and treatment benefits important to patients.

METHODS

Included Response Measures and Criteria

The definition of response combines two elements: the measure being used to assess response, and the criterion, or threshold, used for that measure to determine whether a patient would be defined as a responder or non-responder.

To determine the relevant outcome measures to include in this systematic review, as well as any response criteria associated with these measures, the authors conducted a scoping review of patient endpoints used to assess response in clinical trials and HTA submissions in AD. Inclusion and exclusion criteria for the scoping review are described in Table 1. The outcome measures used as primary endpoints to assess response in phase 3 clinical trials initiated in the last 10 years in adults and/or adolescents with moderate to severe AD were searched for on clinicaltrials.gov (https://www.clinicaltrials.gov, accessed 10 June 2022). Primary endpoints used to assess response were extracted as trials are powered to observe differences in their primary endpoints. HTA submissions were searched using the National Institute for Health and Care Excellence (NICE) and Canadian Agency for Drugs and Technology in Health (CADTH) websites (https://www.nice.org.uk, accessed 15 June 2022 and https://www.cadth.ca, accessed 17 June 2022), and measures used to determine

Table 1 Scoping review criteria for HTA and clinical trials and the response measures and criteria identified

Source	Inclusion criteria	Exclusion criteria	Response measures identified	Response criteria identified
Clinical trials	Moderate or severe AD	Studies in children (aged < 12 years)	EASI	75% improvement in EASI score (EASI-75)
	Studies in adults and/or adolescents (aged ≥ 12 years)			90% improvement in EASI score (EASI-90)
	Phase 3 trials		IGA	IGA score ≤ 1 and a ≥ 2-point improvement
	Study status: recruiting, active not recruiting or completed			IGA score ≤ 1
			PP-NRS	≥ 4-point improvement in PP-NRS score
	Trials started in the last 10 years			
HTA	Moderate or severe AD	Studies in children (aged < 12 years)	EASI	75% improvement in EASI score (EASI-75)
	Studies in adults and/or adolescents (aged ≥ 12 years)	Submissions where committee papers (NICE) or economic reports (CADTH) were not available		90% improvement in EASI score (EASI-90)
			DLQI	50% improvement in EASI score and ≥ 4-point improvement in DLQI score (EASI 50 + DLQI ≥ 4)

Abbreviations: *AD* atopic dermatitis, *CADTH* Canadian Agency for Drugs and Technologies in Health, *DLQI* Dermatology Life Quality Index, *EASI* Eczema Area and Severity Index, *HTA* Health Technology Assessment, *IGA* Investigator's Global Assessment, *NICE* National Institute for Health and Care Excellence, *PP-NRS* Peak Pruritus Numerical Rating Scale

response in the economic model (base case or scenario analyses) were extracted.

The scoping review identified 46 phase 3 trials and five HTAs. The outcome measures used either as primary endpoints in these clinical trials or as definitions of response in the HTA economic model were the clinician-reported Eczema Area and Severity Index (EASI) and Investigator's Global Assessment (IGA) and the patient-reported Dermatology Life Quality Index (DLQI) and Peak Pruritus Numerical Rating Scale (PP-NRS) [2, 8, 9]. The scope of this systematic literature review therefore encompasses these four response measures and the criteria defined from them. A brief description of the included measures is provided in the supplementary material. While a variety of criteria defined from all these measures were used as the primary endpoints in AD clinical trials (Table 1), HTA submissions defined response using either solely EASI or a combined criterion based on the improvement in EASI and DLQI (a 50% improvement in EASI score and ≥ 4 point improvement in DLQI score, i.e. "EASI 50 + DLQI ≥ 4"), which was not used in clinical trials.

Search Strategy

The MEDLINE and Embase databases were searched via ProQuest from database inception to July 21, 2022. The search strategy outlined in Table 2 comprised terms for psychometric validation, disease and symptoms, and included measures. All terms were searched for in titles and abstracts only, and wording variations were captured. The search strategy captured both journal articles and conference abstracts indexed in Embase. All search results were screened by a single reviewer (SJ) using the inclusion/exclusion criteria described in Table 3. All citations first underwent title and abstract screening. The full texts of any articles not excluded at title/abstract level were then evaluated for final inclusion by the same reviewer. Ten percent of the records were double screened by an additional reviewer (HP). Any conflicting decisions were discussed between the two reviewers (SJ and HP) until consensus was reached. The bibliographies of systematic reviews were hand searched to identify studies not captured by the database searches.

Table 2 Search terms and results

Set #	Search terms
S1	(AB,TI("psychometric properties")) OR (AB,TI("psychometric performance")) OR (AB,TI(Valid*)) OR (AB,TI(Reliab*)) OR (AB,TI(Responsive*)) OR (AB,TI(psychometr*)) OR (AB,TI(sensitiv*)) OR (AB,TI("internal consistency"))
S2	(AB,TI(atopic dermatitis)) OR (AB,TI(atopic eczema))
S3	(AB,TI(EASI)) OR (AB,TI(IGA)) OR (AB,TI(DLQI)) OR (AB,TI(PP-NRS)) OR (AB,TI(PP NRS)) OR (AB,TI(Peak Pruritus Numerical Rating Scale)) OR (AB,TI(Eczema Area Severity Index)) OR (AB,TI(Investigator's Global Assessment)) OR (AB,TI(Investigators Global Assessment)) OR (AB,TI(Dermatology Life Quality Index)) OR (AB,TI(Worst itch Numerical Rating Scale))
S4	S1 AND S2 AND S3

Data Extraction

Relevant study and participant characteristics, measures included, methods and results of psychometric testing were extracted using a Microsoft Excel form. Psychometric evidence was extracted in relation to validity (content, convergent, known-group, and structural), reliability (test–retest, inter-rater, and intra-rater) and responsiveness. Additionally, estimates of minimally important differences (MIDs) and minimal important change (MICs) were also extracted where reported. Additional information on how these psychometric properties were defined and tested can be found in Table 4 and the online supplementary material. Any tests of the psychometric performance of a specific response criterion were also extracted. Data extraction was conducted by two reviewers (SJ and SS), with all studies double extracted.

Assessment of Psychometric Performance

Predefined criteria for assessing psychometric performance in relation to each psychometric property were defined in accordance with consensus-based standards for the selection of health measurement instruments (COSMIN) and previous reviews in this area, and are shown in Table 4 [10–12]. Once data extraction for individual studies was performed, the overall evidence was assessed per measure and psychometric property. Overall ratings for each included response measure per psychometric property were defined (Fig. 1).

This article is based on previously conducted studies and does not contain any new studies with human participants or animals performed by any of the authors.

Table 3 Selection criteria for literature

	Inclusion criteria	Exclusion criteria
Population	Persons with AD	Persons with AD not included
	Adults and adolescents (aged ≥ 12 years)	Studies in children aged < 12 years
Intervention	Any or none	N/A
Comparators	Any or none	N/A
Outcomes	Assesses the psychometric performance of one of the following measures:	Measures of interest not included
	EASI	None of the listed elements of psychometric performance tested
	DLQI	
	IGA	
	PP-NRS	
	Psychometric performance of one of the above measures tested in relation to:	
	Content validity	
	Construct validity	
	Structural validity	
	Reliability	
	Responsiveness/sensitivity	
	Minimal (clinically) important differences and/or responder definitions	
	Psychometric performance of any response criteria defined from the above measures tested	
Study design	Clinical or real-world prospective or retrospective studies including chart reviews, database analyses, product or disease registries	Systematic reviews were not included, but relevant included studies were hand searched and included if they met the inclusion criteria
	Cross-sectional survey studies	
Language	Publications in English	Non-English language publications

Abbreviations: *AD* atopic dermatitis, *DLQI* Dermatology Life Quality Index, *EASI* Eczema Area and Severity Index, *IGA* Investigator's Global Assessment, *PP-NRS* Peak Pruritus Numerical Rating Scale

RESULTS

Search Results

Of the 464 unique records retrieved from the MEDLINE and Embase databases via ProQuest, 399 records were excluded at title and abstract screening. Sixty-five papers were reviewed in full, of which 42 were excluded for the reasons outlined in Fig. 2. Twenty-three papers and conference abstracts representing 17 unique studies were included. Six conference abstracts reported on the same study as that reported by

Table 4 Psychometric properties and criteria for good performance

Property	Definition	Criteria for good performance
Validity		
Content validity	The degree to which the items of a PRO measure are an adequate reflection of the construct being measured [11]	Are the included items relevant to the target population, the construct of interest, and the context of use? Are response options appropriate? Is the recall period appropriate? Are there key concepts missing? Are the instructions, items, and response options understood by the target population? Are the items appropriately worded? Do the response options match the question? [12]
Convergent validity	The degree to which scores of an instrument are consistent with hypotheses regarding relationships with scores of other measures [11]	Correlations with similar constructs should be $\geq$ 0.50. Correlations with related but dissimilar constructs should be 0.30–0.50. Correlations with instruments measuring unrelated constructs should be lower than 0.30 [10]
Known-group validity	The degree to which scores of an instrument are consistent with hypotheses regarding differences in scores between relevant groups [11]	Significant differences in scores should be observed across relevant subgroups
Structural validity	Structural validity refers to the degree to which the scores of a measure are an adequate reflection of the dimensionality of the construct being measured [11]	CFA: CFI or TLI $>$ 0.95 or RMSEA $<$ 0.06 or SRMR $<$ 0.08. IRT/Rasch: CFI or TLI $>$ 0.95 or RMSEA $<$ 0.06 or SRMR $<$ 0.08 AND no violation of local independence: (residual correlations among the items after controlling for the dominant factor $<$ 0.20 OR Q3s $<$ 0.37) AND no violation of monotonicity (adequate graphs or item scalability $>$ 0.30) AND adequate model fit (IRT X2 $>$ 0.001 Rasch: infit outfit mean squares $\geq$ 0.50 and $\leq$ 1.50 or Z values > -2 and $<$ 3) [10]
Reliability		
Test–retest reliability	The extent to which scores for patients who have not changed are the same for repeated measurement over time [11]	ICC or weighted kappa $\geq$ 0.70 [10]
Intra-rater reliability	The extent to which scores for patients who have not changed are the same for repeated measurement on different occasions [11]	ICC or weighted kappa $\geq$ 0.70 [10]

Table 4 continued

Property	Definition	Criteria for good performance
Inter-rater reliability	The extent to which scores for patients who have not changed are the same for repeated measurement by different persons on the same occasion [11]	ICC or weighted kappa ≥ 0.70 [10]
Internal consistency	The degree of interrelatedness among items; it is often assessed by Cronbach's alpha [11]	At least low evidence for structural validity and Cronbach ≥ 0.70 for each unidimensional scale [10]
Responsiveness		
Responsiveness	Ability of an instrument to measure a change in health over time [11]	Correlations with changes in similar instruments should be ≥ 0.50. Correlations with changes in similar instruments measuring related but dissimilar constructs should be 0.30–0.50. Correlations with changes in instruments measuring unrelated constructs should be lower than 0.30 AUC should be ≥ 0.70 [10]
MID, MIC or responder definition estimated	MICs provide an estimate of the minimum within-person change over time, which represents a clinically relevant or patient-relevant change. Similarly, an MID represents the minimum clinically relevant difference in score across groups [11]	MICs and MID should be determined using an anchor-based longitudinal approach [10]
Performance of MID, MIC or responder definition tested	The performance of criteria used to determine response should also be tested to ensure that they are effective in distinguishing between those who have and have not experienced a meaningful response	This area is less well defined. Any tests of the performance of response criteria are discussed in this paper

Abbreviations: *AUC* area under the curve, CFA confirmatory factor analysis, *CFI* comparative fit index, *ICC* intraclass correlation coefficient, *IRT* item response theory, *MIC* minimal important change, MID minimally important difference, *PRO* patient-reported outcome, *RMSEA* root mean square error of approximation, *SRMR* standardized root mean square residual, *TLI* Tucker Lewis Index

an included full-text journal article, while four conference abstracts were not covered by a full-text article. Hand-searching systematic reviews resulted in one additional study [13] for a total of 18 unique studies that fulfilled the inclusion criteria for this review.

Characteristics of Included Studies

The key characteristics of the 18 unique studies included in this review are summarized in Table 5. Studies were mainly conducted in the United States of America (USA, *n* = 7), Europe (*n* = 2) and Australia (*n* = 2). Most studies (*n* = 15/18) included adult participants, with the mean age ranging from 30.0 to 52.0 years, whereas one study included only adolescent participants [14]. The age of the participants was not reported in two studies [15, 16]. Where sex was reported (12/18), male participants accounted for 32.6–72.0% of the sample. Sample sizes varied greatly, ranging from 10 to 10,000 participants.

Most studies assessed the validity and reliability of measures ($n = 11$ and $n = 9$, respectively). Six studies also investigated responsiveness, and four studies also estimated MIDs or MICs. Only one study assessed the psychometric performance of a response criterion defined from the measure under investigation [17]. The psychometric results per measure are summarized in Table 6.

EASI

The psychometric performance of the EASI was assessed in seven studies [9, 13, 15, 16, 18–20].

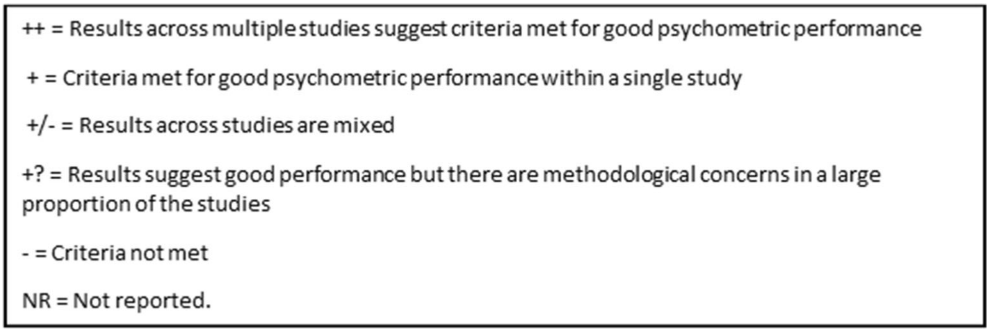
Validity Two studies investigated convergent validity for the EASI [13, 18]. In Bozek 2017, strong correlations between EASI and the objective component of the Scoring Atopic Dermatitis instrument (oSCORAD), the IGA, and patients' assessments of disease severity were reported ($r = 0.66$–0.87) [18]. Shim 2011 reported that the EASI was weakly and insignificantly correlated with a visual analogue scale (VAS) for itch ($r = 0.17$, $P = 0.13$) but moderately and significantly correlated with a VAS for sleep ($r = 0.35$, $P = 0.002$) [13].

Reliability Three studies assessed intra-rater reliability [9, 15, 16, 18]. Bozek 2017 reported good intra-rater reliability for EASI scores (intraclass correlation coefficient [ICC] = 0.71; the two assessments took place on the same day) [18]. Excellent intra-rater reliability was reported in Zhao 2017 and Zhao 2016 (coefficients not reported; the two assessments took place on the same day) [15, 16]. However, an important element in the quality of a study of intra-rater reliability is that the time between administrations should be long enough to prevent easy recall of the initial rating, which is unlikely to be the case with same-day administrations [10].

Inter-rater reliability was assessed in three studies [9, 15, 16]. Zhao 2017 and Zhao 2016 reported good overall inter-rater reliabilities (Zhao 2017: ICC [95% CI] = 0.79 (0.61–0.92); Zhao 2016: ICC in light-skinned patients = 0.85 and ICC in dark-skinned patients = 0.79) [15, 16]. Hanifin 2001 reported good inter-rater reliability using the correlation coefficient of reliability (r-hat > 0.75; the assessments took place on consecutive days). [9]

Responsiveness Only one study assessed the responsiveness of EASI [20]. In Schram 2012, the area under the receiver operating curve (AUC) was 0.67 (95% CI = 0.60–0.76), suggesting poor responsiveness.

MIDs and MICs were estimated in two studies [19, 20]. Schram 2012 reported an MID (anchor: 1-point improvement in IGA) of 6.6 points (standard deviation [SD], 5.9). In this study, the MID varied from 1.0 (IGA from 1 to 0) to 8.6 (IGA from 5 to 4) [20]. In Silverberg



++ = Results across multiple studies suggest criteria met for good psychometric performance

+ = Criteria met for good psychometric performance within a single study

+/- = Results across studies are mixed

+? = Results suggest good performance but there are methodological concerns in a large proportion of the studies

- = Criteria not met

NR = Not reported.

Fig. 1 Definition of overall ratings for each included response measure per psychometric property

2021, 1-point improvements in the Physician's Global Assessment (PGA) and the Validated Investigator's Global Assessment for Atopic Dermatitis (vIGA-AD) were associated with an approximately 50% decrease in EASI score, while a 1-point improvement in the Patient-Reported Global Assessment (PtGA) was associated with a 29.9% decrease in EASI score [19]. One-point improvements in PtGA, PGA, and vIGA-AD scores were associated with approximately 10.9-, 14.0-, and 14.9-point absolute decreases in EASI score. No difference ($P = 0.61$) in the threshold for the EASI-percentage MICs with AD severity was identified, but a significant difference ($P < 0.001$) was observed for the absolute score MIC [19].

IGA

Five studies investigated the psychometric properties of the IGA [16–18, 21, 22].

Validity Convergent validity was assessed in two studies, and known-group validity in one study [18, 21]. In Bozek 2017, strong correlations were observed between the IGA and both the EASI and the oSCORAD ($r = 0.66$–0.80) [18]. In Simpson 2022, strong correlations were observed with the EASI ($r = 0.69$–0.89) and body surface area (BSA; $r = 0.50$–0.75) [21]. Weaker correlations were observed with the Patient-Oriented Eczema Measure (POEM) and the DLQI. These correlations were weak at baseline ($r = 0.30$–0.37) but moderate to strong at week 16 ($r = 0.43$–0.65).

In Simpson 2022, known-group validity was confirmed by comparing vIGA-AD to EASI and Patient Global Impression of Severity—Atopic Dermatitis (PGI-S-AD) severity groups [21]. Patients with a vIGA-AD = 4 were more likely to have worse disease severity on either the EASI or PGI-S-AD compared with patients with a vIGA-AD = 3 ($P < 0.01$).

Reliability One study investigated the test–retest reliability of the vIGA-AD between baseline and week 1 and between weeks 4 and 8

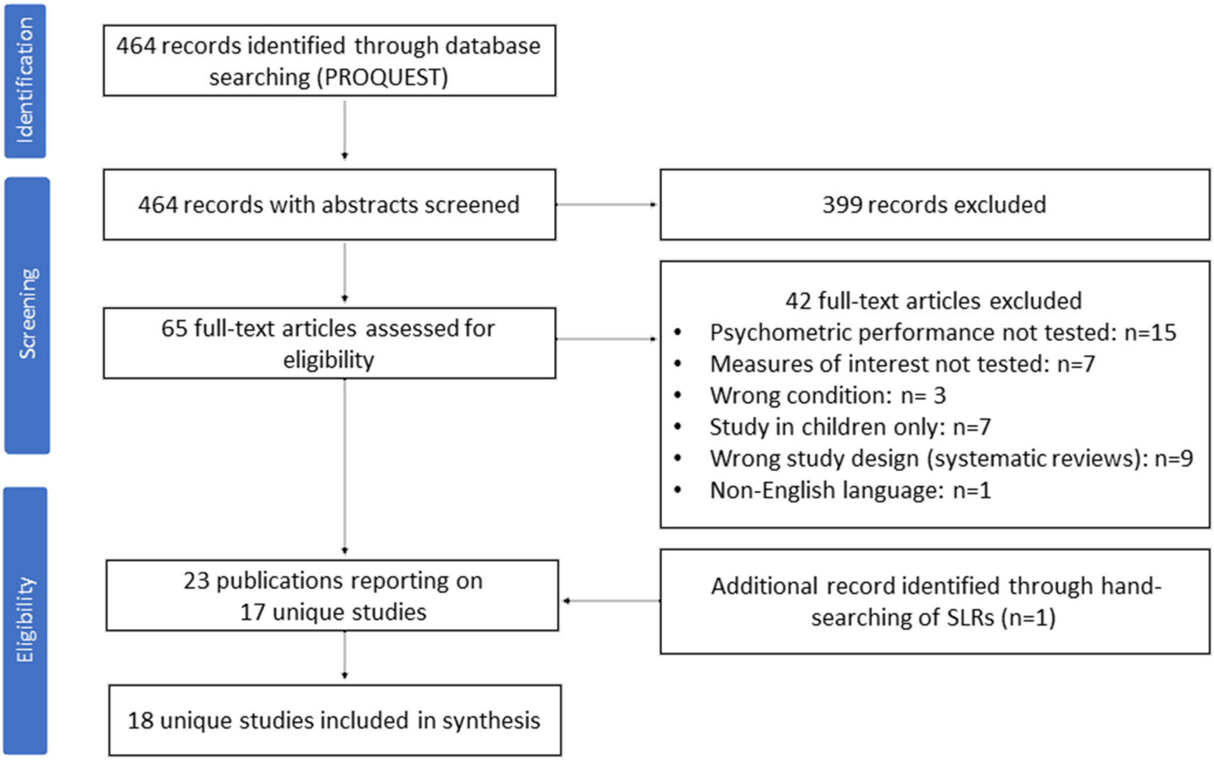

Identification
Screening
Eligibility
464 records identified through database searching (PROQUEST)
464 records with abstracts screened
399 records excluded
65 full-text articles assessed for eligibility
42 full-text articles excluded
• Psychometric performance not tested: n=15
• Measures of interest not tested: n=7
• Wrong condition: n= 3
• Study in children only: n=7
• Wrong study design (systematic reviews): n=9
• Non-English language: n=1
23 publications reporting on 17 unique studies
Additional record identified through hand-searching of SLRs (n=1)
18 unique studies included in synthesis

Fig. 2 PRISMA flow diagram

Table 5 Included study characteristics

First author, year, Ref.	Country	Study design	Number of participants	Age	Male (%)	Measure tested	Other measures included	Psychometric properties evaluated
Bożek 2017 [18]	Poland	NR	10	Mean (SD): 34.9 (12.6)	60.0%	EASI, IGA	oSCORAD	Convergent validity, intra-rater reliability, inter-rater reliability
Hanifin 2001 [9]*	USA	NR	10	Mean = 43.0 years	NR	EASI	NA	Inter-rater reliability and intra-rater reliability
Herd 1997 [25]	UK	Cross-sectional study	10,000	Range: 16–53 years	NR	DLQI	PGI	Convergent validity
Holm 2005 [26]**	Denmark	Case–control study	Cases:101 Controls: 30	NR (66 adults, 35 children with AD and 23 adult and 7 child healthy controls)	NR	DLQI	CDLQI, SF-36, VAS (severity and pruritus), SCORAD	Convergent validity, known-group validity
Patel 2019 [23]	USA	Prospective dermatology practice-based study	340	Mean (SD): 42.8 (16.5)	32.6%	DLQI	POEM, NRS-itch, SCORAD, ItchyQoL, 5D-itch, EASI	Content validity, convergent validity, known-group validity, measurement invariance, internal consistency, responsiveness

Table 5 continued

First author, year, Ref.	Country	Study design	Number of participants	Age	Male (%)	Measure tested	Other measures included	Psychometric properties evaluated
Schram 2012 [20]	NR	Randomized controlled trials	143	Trial 1 Mean age: 40.0 years Trial 2 Mean age: 30.0 years Trial 3 Mean age (adults): 30.9 years	Equally distributed across all trials	EASI	SCORAD, POEM	Responsiveness, MID
Shim 2011 [13]	Korea	NR	83	Mean (SD): 20.6 (10.2)	64.0%	EASI	VAS_{itch} VAS_{sleep}	Convergent validity
Silverberg 2019 [24]	USA	Cross-sectional population-based study	602	Mean (SD): 52.0 (16.3)	46.4%	DLQI	SF-12, SF-6D, PO-SCORAD, PO-SCORAD-itch, PO-SCORAD-sleep, NRS (pain), POEM	Convergent validity, known-group validity, measurement invariance, internal consistency
Silverberg 2019 [17]	USA	Randomized controlled trials	892	Mean (IQR): dupilumab 36.0 (27.0, 47.0); placebo 37.0 (26.0, 49.0)	Dupilumab 58.4%, placebo 54.3%	IGA	EASI, PP-NRS, EQ-5D-3L, BSA, POEM, DLQI	Performance of IGA $\leq$ 1 response criterion
Silverberg 2021 [19]	USA	Prospective observational study	826	Mean (SD): 42.6 (19.3)	47.5%	EASI	POEM, SCORAD, VAS (itch), PGA	MIC

Table 5 continued

First author, year, Ref.	Country	Study design	Number of participants	Age	Male (%)	Measure tested	Other measures included	Psychometric properties evaluated
Simpson 2020 [22]	EU, Japan, Canada, USA	Web-based survey	20	NR (photographic survey 1: 3 pediatric, 17 adult, 15 adolescent/adult; survey 2: 0 pediatric, 10 adult, 15 adolescent/adult)	NA	vIGA-AD	NA	Inter-rater reliability, intra-rater reliability
Simpson 2022 [21]	NR	Pooled RCT data	BREEZE-AD1 = 624; BREEZE-AD2 = 615; BREEZE-AD5 = 440	BREEZE-AD1 = 35.6 (12.8); BREEZE-AD2 = 34.7 (2.8); BREEZE-AD5 = 39.5 (16.1)	62.7%; 62.0%; 50.9%	vIGA-AD	PGI-S-AD EASI BSA POEM DLQI	Convergent validity, known-group validity, test–retest reliability, responsiveness, MID
Sun 2020 [27]	NR	Observational study	570	Mean (SD): 39.1 (16.4)	48.6%	DLQI	NA	Structural validity, internal consistency
Yosipovitch 2018 [14]	USA	Mixed methods (qualitative study and clinical trial)	Qualitative study = 13 patients Clinical trial = 250 patients	Age range: 12–17 years old	NR	PP-NRS	SCORAD, CDLQI, PCS, PGADS, EASI, IGA	Content validity, convergent validity, known-group validity, test–retest reliability, responsiveness

Table 5 continued

First author, year, Ref.	Country	Study design	Number of participants	Age	Male (%)	Measure tested	Other measures included	Psychometric properties evaluated
Yosipovitch 2019 [2]	USA	Mixed methods (qualitative study and clinical trial)	Interviews = 14 Clinical trial: Exploratory analysis using phase IIb data = 379 Confirmatory analysis using pooled phase III data = 1379	Interviews Mean (SD): 40.1 (15.2) Clinical trial Phase IIB mean (SD): 37.0 (12.2) Phase III mean (SD): 38.3 (14.3)	Interviews 35.7% Clinical trial Phase IIb 61.7% Phase III: 57.9%	PP-NRS	PCS, Average Pruritus NRS, itch VAS from SCORAD, itch item from DLQI, PGADS, EASI, IGA	Content validity, convergent validity, known-group validity, test–retest reliability, responsiveness, MID
Yosipovitch 2022 [28]	NR	Mixed methods (qualitative study and clinical trial)	Interviews = 18 Clinical trial = NR	Interviews: Adults: mean (SD) = 30.4 (12.9) Adolescents: mean (SD) = 13.0 (1.0) Clinical trial: Mean: 39.0 years	Interviews: Adults: 27.0% Adolescents: 50.0% Clinical trial: 41.0%	PP-NRS	GAC-AD, IGA	Content validity, convergent validity, test–retest reliability, Responsiveness, MID
Zhao 2016 [16]	Australia	NR	24	NR	NR	EASI	oSCORAD, POEM, PGA	Inter-rater reliability, intra-rater reliability

Table 5 continued

First author, year, Ref.	Country	Study design	Number of participants	Age	Male (%)	Measure tested	Other measures included	Psychometric properties evaluated
Zhao 2017 [15]***	Australia	Prospective study	25	NR (adult and pediatric patients)	72.0%	EASI, IGA	POEM, oSCORAD	Inter-rater reliability, intra-rater reliability

Abbreviations: *AD* atopic dermatitis, *AI-NRS* average itch numeric rating scale, *BSA* body surface area, *CDLQI* Children's Dermatology Life Quality Index, *DLQI* Dermatology Life Quality Index, *EASI* Eczema Area and Severity Index, *EU* European Union, *GAC-AD* Global Assessment of Change–Atopic Dermatitis, *HADS* Hospital Anxiety and Depression Scale, *IGA* Investigator Global Assessment, *ItchyQoL* Itchy Quality of Life; *IQR* interquartile range, *MIC* minimal important change, *MID* minimal important difference, *NA* not applicable, *NR* not reported, *NRS* numerical rating scale, *oSCORAD* Objective Scoring Atopic Dermatitis, *PAS* psychogeriatric assessment scales, *PCS* Pruritus Categorical Scale, *PGA* Physician's Global Assessment; *PGADS* Patient Global Assessment of Disease Status; *PGI* Patient Global Impression; *PGI-S-AD* Patient Global Impression of Symptoms–Atopic Dermatitis, *PNRS* Pruritus Numerical Rating Scale, *PP-NRS* Peak Pruritus Numerical Rating Scale, *POEM* Patient-Oriented Eczema Measure, *PO-SCORAD* Patient-Oriented Scoring Atopic Dermatitis, *SCORAD* Scoring Atopic Dermatitis, *sd* standard deviation, *SF-6D* Short-Form Six-Dimension, *SF-12* 12-Item Short Form Survey, *SF-36* The Short Form (36) Health Survey, *UK* United Kingdom, *USA* United States of America, *vIGA-AD* Validated Investigator Global Assessment for Atopic Dermatitis

*Study reported results from two cohorts: cohort 1 (patients aged 8 years and above) and cohort 2 (patients aged 0 to 7 years). Only results from cohort 1 are reported in this review. **Participants aged < 16 completed the CDLQI instead of the DLQI. Only results from the DLQI were reported. ***Inter-rater intra-class correlation coefficients were reported for adult patients only

Table 6 Summary of psychometric results per measure

	EASI		**IGA**		**DLQI**		**PP-NRS**	
	No. of studies	**Results**	**No. of studies**	**Results**	**No. of studies**	**Results**	**No. of studies**	**Results**
Validity								
Content validity	NR	NR	NR	NR	1	+	3	+ +
Convergent validity	2	±	2	+ +	4	+ +	3	+ +
Known-group validity	NR	NR	1	+	3	+ +	2	+ +
Structural validity	NR	NR	NR	NR	1	+	NR	NR
Reliability								
Test–retest reliability	NR	NR	1	±	NR	NR	3	+ ?
Intra-rater reliability	3	+ ?	2	±	NR	NR	NR	NR
Inter-rater reliability	3	+ +	2	+ +	NR	NR	NR	NR
Internal consistency	NR	NR	NR	NR	3	+ +	NR	NR
Responsiveness								
Responsiveness	1	–	1	+	1	+	3	+ +
MID/MIC/responder definition estimated	2	+	1	+	NR	NR	1	+
Performance of MID/MIC/responder definition tested	NR	NR	1	–	NR	NR	NR	NR

+ + Results across multiple studies suggest that the criteria for good psychometric performance were met; + results from a single study suggest that the criteria for good psychometric performance were met; ± results across studies are mixed; + ? results suggest good performance, but there are methodological concerns in a large proportion of the studies (e.g. very small sample sizes for quantitative studies or issues with the statistical methods used); — criteria not met; *NR* not reported

Abbreviations: *DLQI* Dermatology Life Quality Index, *EASI* Eczema Area and Severity Index, *IGA* Investigator Global Assessment, *MI* minimal important change, *MID* minimal important difference, *PP-NRS* Peak Pruritus Numerical Rating Scale

[21]. When stability was defined using patients reporting no change on the PGI-S-AD, weighted kappas across trials ranged from 0.52 to 0.64, suggesting poor reliability. When using an EASI change below the MID of 6.6 to define stability, weighted kappas ranged from 0.66 to 0.78 [21]. These results suggest borderline reliability according to COSMIN criteria of kappa ≥ 0.7 [10].

Intra-rater reliability was tested in two studies [18, 22]. Bożek 2017 reported an ICC = 0.54 ± 0.28 (the assessments took place on consecutive days) [18], while Simpson 2020 found higher intra-rater reliability, with ICCs > 0.8 between same-day ratings and ratings of the same photograph 5 months apart [22].

Two studies assessed inter-rater reliability. Zhao 2016 reported an ICC (95% CI) of 0.77 (0.58–0.91) in adult patients [16]. Simpson 2020 found very good ICCs and weighted kappa coefficients (0.82–0.89) indicating good inter-rater reliability [22].

Responsiveness One study tested the responsiveness of the vIGA-AD and estimated MIC thresholds [21]. This study reported good responsiveness, as the magnitude of improvement in vIGA-AD scores increased with greater improvement in EASI scores. In the same study, MICs were estimated using anchor-based methods. Overall, the clinical threshold was − 1.00 for minimal meaningful change, − 1.25 to − 1.50 for moderate change, and − 1.75 to − 2.00 for large change. Distribution-based methods gave estimates of − 0.25 (0.5 baseline SD) and − 0.65 (minimal detectable change with 95% confidence). [21]

Silverberg 2019 investigated the performance of the IGA response criterion of IGA ≤ 1 in AD patients [17]. This study found that people defined as non-responders on the IGA (IGA > 1) had clinically meaningful improvements in EASI, PP-NRS, EQ-5D-3L, DLQI, POEM, and BSA scores, suggesting that the IGA response criterion of IGA ≤ 1 may be too restrictive and may not account for the meaningful benefit from itch relief and decreases in the extent and severity of AD lesions and in overall quality of life [17].

DLQI

Five studies investigated the psychometric properties of the DLQI [23–27].

Validity One study reported good content validity for the DLQI in patients with AD, with 92% of patients reporting that the DLQI covered all the issues most relevant to them in relation to AD [23]. The four participants who reported that the DLQI did not assess their most important issues identified sleep disturbance as the most important symptom. Some participants reported that items concerning sports and sexual activity were not important to them, as they did not participate in these activities, but no participant considered any items conceptually irrelevant.

Convergent validity was assessed in four studies using clinical trial data in AD [23–26]. Silverberg 2019 reported strong correlations between DLQI and the Patient-Oriented SCORAD (PO-SCORAD; $r = 0.71$) and the POEM ($r = 0.62$), and moderate correlations with the PO-SCORAD itch subscore (PO-SCORAD-itch; $r = 0.48$) and the Numerical Rating Scale (NRS) for pain (NRS-pain; $r = 0.43$, $P < 0.001$ for all) [24]. Patel 2019 reported strong correlations with the Itchy Quality of Life (ItchyQOL), the 5-D Itch Scale (5-D itch), the NRS for average itch (NRS-itch), the POEM, and the SCORAD ($r = 0.55$–0.79), and a moderate correlation with the EASI (0.44) [23]. In Holm 2005, the DLQI correlated strongly with a pruritus VAS (PRU-VAS), a patient disease severity VAS (PTVAS), and an investigator overall assessment VAS (INVAS; $r = 0.62$–0.82), moderately with the 36-Item Short Form Health Survey mental component score (SF-36 MCS; $r = -0.46$) and weakly with the 36-Item Short Form Health Survey physical component score (SF-36 PCS; $r = -0.27$) [26]. Herd 1999 found a strong correlation between DLQI and Patient-Generated Index (PGI; $r = 0.52$, $P < 0.001$) and moderate correlations with health service costs ($r = 0.47$) and total costs ($r = 0.34$). [25]

Known-group validity was assessed in three studies. Patel 2019 found significant stepwise increases in DLQI score at each level of severity measured by the POEM, NRS-itch, EASI, and SCORAD instruments ($P < 0.0001$) [23]. Similarly, Holm 2005 found that DLQI scores were significantly associated with SCORAD severity groups ($P < 0.0001$) [26]. In Silverberg 2019, DLQI scores increased significantly with each increasing level of severity on self-reported global AD severity, the POEM, the PO-SCORAD, the PO-SCORAD-itch, the PO-SCORAD sleep subscore (PO-SCORAD-sleep), and the NRS-pain (analysis of variance, $P < 0.0001$ for all). AUC analysis showed that the DLQI was excellent at distinguishing between severe versus mild AD and good at distinguishing moderate versus mild or severe versus moderate AD, outperforming the 12-Item Short Form Health Survey (SF-12) [24].

One study assessed the structural validity of the DLQI. Sun 2020 used an explanatory factor analysis and found two factors: factor 1 (items 1–7) assessed personal life, and factor 2 (items 8–10) assessed social factors and treatments (eigenvalues = 5.00 and 1.13, respectively) [27]. A bifactor model indicated a good fit (root mean square error of approximation [RMSEA] = 0.046; comparative fit index [CFI] = 0.988), with standardized factor loadings on the general factor of

0.42–0.82. The global factor explained 93.5% of the common variance, whereas the specific factors explained 0.4% and 6.1%, indicating sufficient unidimensionality.

Reliability Good internal consistency was confirmed in three studies reporting Cronbach's alphas of 0.94–0.89 for the DLQI [24, 27].

Responsiveness The DLQI's responsiveness was assessed in one study. Patel 2019 observed medium standardized effect sizes (Cohen's D) in the anticipated direction in DLQI scores for those participants who experienced a change in POEM score ≥ 3.4 points (a previously determined MID), suggesting good responsiveness (Cohen's D = |0.65–0.72|) [23].

PP-NRS

Three studies investigated the performance of the PP-NRS in patients with AD [2, 14, 28].

Validity Three studies reported good content validity in qualitative interviews with AD patients, with the PP-NRS found to be relevant, appropriate, well understood, and consistently interpreted [2, 14, 28]. All patients across two studies reported itch as a core concept/symptom of their AD [2, 28], while 93% reported at least one meaningful consequence of itch, including embarrassment/self-consciousness, bleeding, problems with concentration, and sleep disturbance [2]. Participants considered both worst and average itch over the past 24 h to be important and comprehensive in assessing itch, but found worst itch easier to rate and more important to improve with treatment [2].

Convergent validity was assessed in three studies using clinical trial data [2, 14, 28]. Moderate to strong correlations between PP-NRS and other patient-reported outcomes (PROs) were found in Yosipovitch 2018 (PROs: SCORAD-itch, Children's Dermatology Life Quality Index–itch [CDLQI-itch], Pruritus Categorical Scale [PCS], and Patient Global Assessment of Disease [PGAD]) and Yosipovitch 2022 (PROs: DLQI, DLQI-itch, POEM, and POEM-itch) at baseline ($r = 0.41$–0.73) and strong correlations at week 16 ($r = 0.64$–0.83) [14, 28]. Both studies found weaker correlations with clinician-reported outcomes (Yosipovitch 2018: EASI and IGA; Yosipovitch 2022: EASI, IGA and BSA). These correlations were weak to moderate at baseline ($r = 0.20$–0.31) and moderate to strong at week 16 ($r = 0.43$–0.53). In Yosipovitch 2019, strong correlations were observed with the PCS, the DLQI-itch, and the SCORAD-itch VAS ($r = 0.61$–0.77), and weak correlations were observed with the EASI and the IGA ($r = 0.09$–0.24) at baseline [2].

Known-group validity was assessed in two studies [2, 14]. Baseline and week-16 PP-NRS scores differed predictably across CDLQI and PCS levels (F-statistic; all $P < 0.0001$) in Yosipovitch 2018 [14]. Similarly, scores varied significantly between categories, reflecting no/mild itch/symptoms versus severe itch/symptoms on the PCS, the DLQI, and the Patient Global Assessment of Disease Status (PGADS; $P < 0.0001$ for all comparisons) in Yosipovitch 2019 [2].

Reliability Test–retest reliability was confirmed in three studies [2, 14, 28]. Yosipovitch 2018 reported that coefficients between baseline to week 2 and week 15 to week 16 exceeded the recommended threshold of 0.7 (the exact test and coefficients are unclear) [14]. Yosipovitch 2019 and 2022 reported ICCs ≥ 0.89, indicating very good test–retest reliability (2019: assessments at weeks 15 and 16; 2022: assessments at weeks 12 and 16), although only Yosipovitch 2022 reported testing this exclusively in participants defined as stable across test–retest periods (using IGA score) [2, 28].

Responsiveness Responsiveness was assessed in three studies using both correlations of mean changes with similar measures and effect sizes [2, 14, 28]. Yosipovitch 2018 reported moderate to strong correlations in change scores with similar PROs ($r = 0.40$–0.68) and significant patterns of mean change across PGAD levels (F-statistic, 23.7; $P < 0.0001$) [14]. Yosipovitch 2019 reported strong correlations of change scores with PCS, DLQI-itch, and SCORAD-itch VAS ($r = 0.64$–0.77) and moderate to strong correlations with changes in EASI and IGA scores ($r = 0.46$–0.50) [2]. Large effect size estimates of change were reported in all three studies [2].

Yosipovitch 2022 reported an MIC of 3 points, estimated with anchor-based methods, using the IGA and the Global Assessment of Change–Atopic Dermatitis (GAC-AD) as anchors [28]. In the qualitative phase, most participants stated that a 2-point ($n = 6$) or 3-point ($n = 10$) decrease in PP-NRS indicated meaningful improvement. Yosipovitch 2019 reported MIC estimates based on clinician-reported and patient-reported anchors (EASI, IGA and PCS) ranging between 2.2- and 4.2-point improvements [2].

DISCUSSION

The scoping review of clinical trials and HTA submissions uncovered important findings about how response is measured in AD. EASI, IGA, DLQI and PP-NRS were used to define response using various criteria. There was little consistency in the assessment of response, both across clinical trials and between trials and HTAs. While a variety of criteria defined from these measures were used as the primary endpoint in clinical trials, HTA submissions defined response using either EASI score alone or a combined criterion based on improvement in EASI and DLQI scores, something not used in clinical trials. The identified lack of consistency in the assessment of response observed in the scoping review makes it difficult for clinicians, regulators, and payers to directly compare the efficacies of different treatments to make optimal treatment and resource allocation decisions. Psychometric evidence on the performance of response measures and criteria should be used to guide decisions on which are most appropriate for use to facilitate consistency.

While this review identified some evidence on the psychometric performance of the measures being used to assess treatment response in AD, important gaps in the evidence were revealed. Content validity was only assessed for the patient-reported DLQI and PP-NRS. No assessments of the content validity of the clinician-reported EASI and IGA were identified. Content validity is arguably the most important psychometric property, as it determines whether the measure covers what is important to patients without including irrelevant items, and is understood as intended [10, 12]. Content validation of the PP-NRS found that patients reported itch as a core symptom that had an important impact on their daily life and was a priority for treatment. While the EASI and the IGA are established measures of clinical severity, results of several studies called into question their coverage of patient-relevant symptoms. One included study found that the EASI did not correlate with an itch VAS, indicating a lack of coverage of itch [13]. Another study found that people defined as non-responders by the IGA had clinically meaningful improvements in itch, extent and severity of AD lesions, and overall quality of life, concepts not covered by the IGA [17]. It is vital to investigate the content validity of the EASI and the IGA. These measures are frequently used to assess efficacy and response in clinical trials and HTAs, but they may miss key elements of patient-relevant disease impact and treatment benefit, including itch, which leads to an inadequate understanding and estimation of treatment efficacy. This will result in treatment efficacy and cost-effectiveness being undervalued in regulatory and HTA decision-making, decreasing the chances of treatment acceptance and reimbursement. Moreover, response criteria used in clinical trials and by HTA bodies will likely find their way to the prescriber setting, whereby non-responders would not get their treatment reimbursed. In this case, if response assessment does not fully capture patient-relevant benefit, this may hamper patients' access to tailored treatment.

Several studies estimated responder definitions for one of the included measures or reported responsiveness results using a predefined response criterion. Only one study investigated the performance of a predefined response criterion for the IGA [17]. However, no studies were identified that compared the psychometric performance of alternative response criteria to make recommendations on an appropriate and consistent definition of response. Such a definition could inform high-quality evidence synthesis for comparisons of the efficacy and value of different treatments. Such studies are available in other conditions, including rheumatoid and psoriatic arthritis and cancer [7, 29–31]. A good response criterion should capture the symptoms, impact of disease, and elements of treatment benefit that are important to patients, and it should be able to discriminate between patients receiving a meaningful treatment benefit and those who are not [7, 30]. Evidence is required on the comparative performance of the different criteria being used to define response to inform which are able to comprehensively capture patient-relevant treatment benefits and distinguish those patients receiving effective treatment. This evidence would pave the way for the standardization of response assessment, which would enable high-quality estimates of the comparative efficacy of treatments and evidence-based regulatory, HTA and clinical decision-making.

Limitations of This Review

The reliability of the results of the included studies is dependent on the quality of those studies. Some quantitative studies were performed using very small sample sizes. Results for convergent and known-group validity are dependent on the appropriateness of comparator measures and the known groups defined. Not all investigations of test–retest reliability reported an anchor, based on which the population could be assumed to be stable in health over time. Including patients who were not stable over the test–retest period would impact results. As some studies were reported in conference abstracts only, sufficient details to be able to judge how statistical tests were performed were sometimes unavailable. Different versions of IGA scales were used across studies. Studies often failed to provide the exact wording of the IGA used in their study, and therefore it is unclear to what extent results on the IGA from different studies are comparable.

Lastly, this review was focused on the psychometric performance of measures and criteria which have been used to assess response in phase 3 clinical trials and HTAs in AD in the last 10 years. Other measures are available which capture patient-relevant endpoints such as itch, including SCORAD and POEM [32, 33]. Future research could investigate the extent to which such measures would be suitable to assess response as primary outcomes in clinical trials and HTAs in AD.

CONCLUSION

The current landscape of disjointed evidence on the responsiveness of different treatments, with different response measures and criteria used, makes direct comparisons of treatment efficacy nearly impossible for clinicians, regulators, and payers. This impedes evidence-based treatment and sound resource allocation decisions. While content validation of the PP-NRS confirmed the importance of itch as a core symptom and treatment priority in AD, the EASI and IGA lack both coverage of itch and content validation. It is concerning that itch is currently not well covered in response assessments, while there is a patient-relevant instrument available with sound psychometric properties, namely the PP-NRS. Including the PP-NRS in both clinical trials and HTAs will place more emphasis on patient-reported benefit and response. Although response thresholds are estimated in some studies, no studies have compared the psychometric performance of different response criteria to inform which were appropriate to use to compare treatments and to pave the way towards a consistent definition of response across trials and HTA.

ACKNOWLEDGEMENTS

The authors want to thank An Ta for her contributions to the literature review. Her contribution was funded by Galderma.

Medical Writing/Editorial Assistance The authors want to thank Christina DuVernay, an employee of OPEN Health, for her editorial preparation of this manuscript. Her contribution was funded by Galderma.

Authorship All named authors meet the International Committee of Medical Journal Editors (ICMJE) criteria for this article, take full responsibility for the integrity of the work as a whole, and have given their approval for this version to be published.

Author Contributions. All authors were involved in the conceptualization of this study. Hannah Penton, Jorge Puelles, Marieke Heisen and Sayeli Jayade all contributed to the design of the study. Hannah Penton, Sayeli Jayade and Santhani Selveindran managed the acquisition, screening, extraction, analysis and interpretation of the data. Hannah Penton, Sayeli Jayade, Santhani Selveindran and Marieke Heisen drafted the manuscript, which was critically reviews by Christophe Piketty, Liliana Ulianov, Zarif Jabbar-Lopez, Jonathan I. Silverberg and Jorge Puelles. All authors provided approval for the publication of the final version of the manuscript.

Funding. Hannah Penton, Sayeli Jayade, Santhani Selveindran and Marieke Heisen are employees of OPEN Health. OPEN Health received funding from Galderma for the conduct of this study. Christophe Piketty, Liliana Ulianov, Zarif K. Jabbar-Lopez and Jorge Puelles are employees of Galderma. Galderma is funding the journal's Rapid Services Fee.

Declarations

Conflict of Interest. Christophe Piketty, Liliana Ulianov, Zarif K. Jabbar-Lopez and Jorge Puelles are employees of Galderma, which funded this research. Hannah Penton, Sayeli Jayade, Santhani Selveindran and Marieke Heisen are employees of OPEN Health, which received consulting fees from Galderma in connection with this study. Jonathan I. Silverberg has received honoraria as a consultant and/or advisory board member for AbbVie, Alamar, Aldena, Amgen, AObiome, Arcutis, Arena, Asana, Aslan, BioMX, Biosion, Bodewell, Boehringer Ingelheim, Cara, Castle Biosciences, Celgene, Connect Biopharma, Dermavant, Dermira, Dermtech, Eli Lilly, Galderma, GlaxoSmithKline, Incyte, Kiniksa, Leo Pharma, Menlo, Novartis, Optum, Pfizer, RAPT, Regeneron, Sanofi-Genzyme, Shaperon, and Union, and as a speaker for AbbVie, Eli Lilly, Leo Pharma, Pfizer, Regeneron, and Sanofi-Genzyme; his institution received grants from Galderma, Incyte, and Pfizer.

Ethical Approval. This article is based on previously conducted studies and does not contain any new studies with human participants or animals performed by any of the authors.

REFERENCES

1. Nutten S. Atopic dermatitis: global epidemiology and risk factors. Ann Nutr Metab. 2015;66(Suppl 1): 8–16.

2. Yosipovitch G, Reaney M, Mastey V, et al. Peak pruritus numerical rating scale: psychometric validation and responder definition for assessing itch in moderate-to-severe atopic dermatitis. Br J Dermatol. 2019;181(4):761–9.

3. Schmitt J, Csotonyi F, Bauer A, Meurer M. Determinants of treatment goals and satisfaction of patients with atopic eczema. J Dtsch Dermatol Ges. 2008;6(6):458–65.

4. Nusbaum KB, Fleischer S, Fleischer AB. Efficacy of biologics and oral small molecules for atopic dermatitis: a systematic review and meta-analysis. J Dermatol Treat. 2021. https://doi.org/10.1080/09546634.2021.1986204.

5. Schmitt J, Langan S, Williams HC. What are the best outcome measurements for atopic eczema? A systematic review. J Allergy Clin Immunol. 2007;120(6):1389–98.

6. Charman C, Williams H. Outcome measures of disease severity in atopic eczema. Arch Dermatol. 2000;136(6):763–9.

7. Verhoeven AC, Boers M, van Der Linden S. Responsiveness of the core set, response criteria, and utilities in early rheumatoid arthritis. Ann Rheum Dis. 2000;59(12):966–74.

8. Finlay AY, Khan GK. Dermatology Life Quality Index (DLQI)—a simple practical measure for routine clinical use. Clin Exp Dermatol. 1994;19(3):210–6.

9. Hanifin JM, Thurston M, Omoto M, Cherill R, Tofte SJ, Graeber M. The eczema area and severity index (EASI): assessment of reliability in atopic dermatitis. EASI Evaluator Group Exp Dermatol. 2001;10(1):11–8.

10. Prinsen CAC, Mokkink LB, Bouter LM, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018;27(5):1147–57.

11. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010;63(7):737–45.

12. Terwee CB, Prinsen CAC, Chiarotto A, et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. Qual Life Res. 2018;27(5):1159–70.

13. Shim WHPH, Kim HS, Kim SH, Ko HC, Kim MB, Kim DW, Kim BS. Does the EASI score reflect itch severity? Ann Allergy Asthma Immunol. 2011;106:540–1.

14. Yosipovitch G, Guillemin I, Eckert L, et al. Validation of the Peak Pruritus Numerical Rating Scale (NRS) in adolescent moderate-to-severe atopic dermatitis patients for use in clinical trials. 3rd Inflammatory Skin Disease Summit—The Translational Revolution. Exp Dermatol. 2018;27:42.

15. Zhao CY, Hao EY, Oh DD, et al. A comparison study of clinician-rated atopic dermatitis outcome measures for intermediate- to dark-skinned patients. Br J Dermatol. 2017;176(4):985–92.

16. Zhao C, Hao E, Oh D, et al. Validation of a novel grey-scale and atopic dermatitis scores for skin of colour patients—results from a multi-center study. In: 49th Annual Scientific Meeting of the Australasian College of Dermatologists; 2016 May 14–17; Perth, Australia. p. 85.

17. Silverberg JI, Simpson EL, Ardeleanu M, et al. Dupilumab provides important clinical benefits to patients with atopic dermatitis who do not achieve clear or almost clear skin according to the Investigator's Global Assessment: a pooled analysis of data from two phase III trials. Br J Dermatol. 2019;181(1):80–7.

18. Bozek A, Reich A. Assessment of intra- and inter-rater reliability of three methods for measuring atopic dermatitis severity: EASI, objective SCORAD, and IGA. Dermatology. 2017;233(1):16–22.

19. Silverberg JI, Lei D, Yousaf M, et al. What are the best endpoints for eczema area and severity index and scoring atopic dermatitis in clinical practice? a prospective observational study. Br J Dermatol. 2021;184(5):888–95.

20. 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.

21. Simpson EL, Bissonnette R, Paller AS, et al. The Validated Investigator Global Assessment for Atopic Dermatitis (vIGA-AD): a clinical outcome measure for the severity of atopic dermatitis. Br J Dermatol. 2022;187(4):531–8.

22. Simpson E, Bissonnette R, Eichenfield LF, et al. The Validated Investigator Global Assessment for Atopic Dermatitis (vIGA-AD): the development and reliability testing of a novel clinical outcome measurement instrument for the severity of atopic dermatitis. J Am Acad Dermatol. 2020;83(3):839–46.

23. Patel KR, Singam V, Vakharia PP, et al. Measurement properties of three assessments of burden used in atopic dermatitis in adults. Br J Dermatol. 2019;180(5):1083–9.

24. Silverberg JI, Gelfand JM, Margolis DJ, et al. Validation and interpretation of short form 12 and comparison with dermatology life quality index in atopic dermatitis in adults. J Invest Dermatol. 2019;139(10):2090–7.

25. Herd RM, Tidman MJ, Ruta DA, Hunter JA. Measurement of quality of life in atopic dermatitis:

correlation and validation of two different methods. Br J Dermatol. 1997;136(4):502–7.

26. Holm EA, Wulf HC, Stegmann H, Jemec GB. Life quality assessment among patients with atopic eczema. Br J Dermatol. 2006;154(4):719–25.

27. Sun X, Li X, Arenson E, et al. Dimensional structure of dermatology life quality index in patients with atopic dermatitis: bifactor model approach ISPOR Europe 2020 virtual. Value Health. 2020;23:S676–7.

28. Yosipovitch G, Rams A, Baldasaro J, et al. Content validity and assessment of the psychometric properties and score interpretation of a pruritus numeric rating scale in atopic dermatitis. Revolutionizing atopic dermatitis. Br J Dermatol. 2022;186(4):e145.

29. Fransen J, Antoni C, Mease PJ, et al. Performance of response criteria for assessing peripheral arthritis in patients with psoriatic arthritis: analysis of data from randomised controlled trials of two tumour necrosis factor inhibitors. Ann Rheum Dis. 2006;65(10):1373–8.

30. van Gestel AM, Prevoo MLL, et al. Development and validation of the European League Against Rheumatism response criteria for rheumatoid arthritis. Comparison with the preliminary American College of Rheumatology and the World Health Organization/International League Against Rheumatism Criteria. Arthritis Rheum. 1996;39(1): 34–40.

31. Castello A, Rossi S, Toschi L, Lopci E. Comparison of metabolic and morphological response criteria for early prediction of response and survival in NSCLC patients treated with anti-PD-1/PD-L1. Front Oncol. 2020;10:1090.

32. European Task Force on Atopic Dermatitis. Severity scoring of atopic dermatitis: the SCORAD index. 1993 Consensus Report of the European Task Force on Atopic Dermatitis. Dermatology. 1993;186(1): 23–31.

33. Charman CR, Venn AJ, Williams HC. The patient-oriented eczema measure: development and initial validation of a new tool for measuring atopic eczema severity from the patients' perspective. Arch Dermatol. 2004;140(12):1513–9.