

Psychometric validation of the physician global assessment scale for assessing severity of psoriasis disease activity

J. C. Cappelleri · A. G. Bushmakina ·
J. Harness · C. Mamolo

Accepted: 25 February 2013 / Published online: 9 March 2013
© Springer Science+Business Media Dordrecht 2013

Abstract

Purpose The Physician Global Assessment (PGA) is a key measure of psoriasis frequently used in clinical trials. A psychometric validation of a three-item (erythema, induration, and scaling) PGA scale was performed using Phase 2 data.

Methods Confirmatory factor analysis (CFA) tested the PGA measurement model and appropriateness of equal weighting of the items. PGA test–retest reliability was assessed by estimating the intraclass correlation coefficient (ICC). Internal consistency reliability was gauged by calculating Cronbach’s coefficient α ($CC\alpha$). Clinically important difference (CID) was defined using the repeated measures model to estimate the relationship between PGA and Patient Global Assessment (PtGA). Known-group, convergent, and divergent validity for the PGA were also assessed.

Results 197 patients with psoriasis were randomized to tofacitinib 2, 5, 15 mg twice daily, or placebo. CFA demonstrated that the PGA measurement model fitted the data using equal weighting of the PGA items. The PGA scale demonstrated good test–retest reliability ($ICC > 0.7$) and internal consistency reliability ($CC\alpha > 0.8$). The CID for PGA was estimated at 0.52 (95 % confidence interval:

0.47, 0.56). A robust monotonic relationship between PGA and Psoriasis Area and Severity Index (PASI) data substantiated known-group validity. Relatively high correlations of PGA with PASI and PtGA data (all correlations >0.5 except at baseline) supported convergent validity; relatively low correlations of PGA with the Pain/Discomfort Assessment and the Ocular Comfort Index supported divergent validity.

Conclusions The three-item PGA scale has sound psychometric properties with respect to reliability and validity, with equal weighting of the items being appropriate.

Keywords Physician Global Assessment · Psoriasis · Validation

Introduction

The Physician Global Assessment (PGA) is a key measure of psoriasis severity frequently used in clinical trials as an efficacy endpoint. More widely used is the Psoriasis Area and Severity Index (PASI), which combines the assessment of the severity of lesions with the body area affected [1]. In clinical trials, both tools are often used as efficacy endpoints since neither has a distinct advantage over the other [2]; indeed, this approach is strongly recommended in the European Medicines Agency Committee for Medicinal Products for Human Use (CHMP) guidance on psoriasis [3]. However, there remains some disagreement as to the reliability and variability of the PASI and PGA assessments [4–8].

The PASI was developed in 1978 by Fredriksson and Pettersson [1] to assess the outcome of treatment in psoriasis by quantifying lesion severity and percentage of body area affected. The plaque characteristics erythema, induration, and scaling are assessed separately for four body

J. C. Cappelleri (✉) · A. G. Bushmakina
Pfizer Inc, 445 Eastern Point Road, MS 8260-2502, Groton,
CT 06340, USA
e-mail: joseph.c.cappelleri@pfizer.com

J. Harness
Novartis Pharma AG, Basel, Switzerland

C. Mamolo
Pfizer Inc, 445 Eastern Point Road, MS 8260-2510, Groton,
CT 06340, USA

regions: head, upper limbs, trunk, and lower limbs. Severity is rated on a 0–4 scale (0, no involvement; up to 4 for very severe involvement) for each symptom, which is summed by body region before then being multiplied by a weighting factor to account for the percent of affected body surface area within each body region (0, no area affected; up to 6 for $\geq 90\%$ area affected). The individual body region scores are then multiplied by another weighting factor reflecting their respective proportion of total body surface area (head 0.1; upper limbs 0.2; trunk 0.3; lower limbs 0.4). The sum of the resultant values for each of the 4 body region provides the PASI. The PASI score can range from 0 (representing no psoriasis) to 72 (vast majority of body covered by severe psoriasis plaques).

In review of assessment tools in randomized-controlled psoriasis trials published between 1977 and 2000 [9], of those trials including a clinical scoring system, 48.5 % (83/171) used the PASI. Although the PASI scoring system is the most popular endpoint in clinical trials for psoriasis to determine disease severity, it is rarely used in clinical practice due to the complex assessment and scoring required. Other limitations of the PASI include that it lacks sensitivity at the lower range of the scale (i.e., for assessing milder disease, especially with limited body surface area), with only the lower half of the scale used in practice (e.g., very severe psoriasis will score in the 40–50 PASI range, with scores of 50–72 on the PASI very rarely seen). PASI scoring is not linear, making it difficult to interpret, and inter-rater reliability can be low without thorough training of the raters [7, 10, 11].

While the PASI scoring system is standardized, there is no one standard PGA scale in use. PGA scales differ in the number of symptoms ('items') assessed, as well as the number of categories within each scale, and the description of each category. Most common is a 'static' PGA scale that refers to the disease at the current point in time; however, 'dynamic' PGA scales are also used, which assess the change in disease severity from a reference point. The European Medicines Agency Committee for Medicinal Products for Human Use (CHMP) guidance on psoriasis issued in 2004 gives an example of a static, single-item PGA scale that has 6 categories (while noting that 7 category scales are also used, and that the wording for PGA scales varies). This example of a single-item PGA scales ranges from 'clear' (no signs of psoriasis [post-inflammatory hyperpigmentation may be present]) to 'severe' (very marked plaque elevation, scaling, and/or erythema). As this example shows, the category descriptors mention the three main symptoms (erythema, induration, scaling); all of which may or may not be present in varying degrees [3]. This variability can lead to differences among raters in how to interpret and assign a single-item category when the symptom severities are not all aligned (e.g., such as in a situation with mild plaque elevation, mild scaling yet with moderate erythema).

The PGA scale, unlike the PASI, does not take into account the amount of body surface area affected by plaques. Rather than considering a patient as split into four discrete body regions for disease grading, instead, somewhat abstractly, the PGA considers the whole patient. The PGA is also a patient-centered response metric including categories for being 'clear' and 'almost clear' [8, 12]. The PGA categories are easy to interpret, and the entire range of the scale is used [6]. For the tofacitinib psoriasis development program, a three-item (erythema, induration and scaling) PGA was developed, with each item assessed according to a separate and item-specific five-category scale. Assessing three different items may provide better precision and reliability relative to a single item, since a multi-item scale is generally more precise and reliable than a single-item scale of the same concept [13], especially with a complex entity like psoriasis severity. Differing symptom severities for the three items can also be clearly reflected. The PGA score is the mean of the three separate item scores rounded to the nearest whole number (Fig. 1a).

In this paper, we evaluated the psychometric properties of the three-item PGA, by analyzing its reliability and validity, using data from a Phase 2 study investigating tofacitinib monotherapy for the treatment of patients with moderate-to-severe chronic plaque psoriasis [14].

Methods

Patients

In this randomized, double-blind, parallel-group, placebo-controlled, multicenter study, tofacitinib was investigated for the treatment of adult patients with moderate-to-severe chronic plaque psoriasis. Full details of the trial have been described elsewhere [14]. Briefly, patients were randomized to receive oral tofacitinib 2, 5, or 15 mg twice daily (BID) or placebo for 12 weeks (ClinicalTrials.gov ID: NCT00678210). This study was performed in compliance with the International Conference on Harmonization Good Clinical Practice Guidelines; all patients provided written informed consent, and institutional review boards or ethics committees approved the protocol before the study started.

Psoriasis assessment

Measures of psoriasis severity were evaluated throughout the study by the patient and by trained professionals. Where possible, the same evaluator assessed all efficacy endpoints for any individual patient throughout the study. Assessment for the following outcomes was collected at multiple times (as noted below) and, if applicable, at an early termination visit.

Fig. 1 **a** Physician's Global Assessment (PGA). **b** Patient's Global Assessment (PtGA)**(a) PGA**

Erythema, induration and scaling are rated separately across all psoriatic lesions. Severity rating scores are summed and the average taken, with the average rounded to the nearest integer to determine the PGA score.

ERYTHEMA (E)
Determine erythema (averaged over the whole body)
☐ 0 = no evidence of erythema (post-inflammatory hyperpigmentation and/or hypopigmentation may be present)
☐ 1 = light pink
☐ 2 = light red
☐ 3 = red
☐ 4 = dark, deep red

INDURATION (I)
Determine induration (averaged over the whole body)
☐ 0 = no evidence of plaque elevation
☐ 1 = barely palpable
☐ 2 = slight but definite elevation, indistinct edge
☐ 3 = elevated with distinct edges
☐ 4 = marked plaque elevation, hard/sharp borders

SCALING (S)
Determine scaling (averaged over the whole body)
☐ 0 = no evidence of scaling
☐ 1 = occasional fine scale
☐ 2 = fine scale predominates
☐ 3 = coarse scale predominates
☐ 4 = thick, coarse scale predominates

Add E + I + S = / 3 = (Total Average)

PHYSICIAN GLOBAL ASSESSMENT - based upon above Total Average
☐ 0 = Clear - cleared, except for any residual discoloration
☐ 1 = Almost clear - majority of lesions have individual scores for E + I + S / 3 that average 1
☐ 2 = Mild - majority of lesions have individual scores for E + I + S / 3 that average 2
☐ 3 = Moderate - majority of lesions have individual scores for E + I + S / 3 that average 3
☐ 4 = Severe - majority of lesions have individual scores for E + I + S / 3 that average 4

Note: Total average is rounded to the nearest whole number score, eg if total ≤ 2.49 , score = 2; if total ≥ 2.50 , score = 3

(b) PtGA

Overall, how would you describe your psoriasis right now?

Choose only ONE response.

- My psoriasis is: ☐ Severe
☐ Moderate
☐ Mild
☐ Almost Clear
☐ Clear (no psoriasis)

PGA

The PGA scale ranged from 0, ‘Clear’ to 4, ‘Severe’ (Fig. 1a). PGA was assessed at Baseline and at Weeks 2, 4, 6, 8, 12 (end of treatment), 14 and 16 (2- and 4-week follow up), and at early termination. Three items (erythema, induration, and scaling) were used to define the latent (or underlying) PGA construct of interest. Items were rated separately across all psoriatic lesions and scored from 0 to 4 based on the appropriate morphologic descriptors. Severity rating scores for each item were summed and the average taken; the average was rounded to the nearest integer to determine the PGA score [3, 15]. The PGA reflects a global, overall patient consideration of the erythema, induration, and scaling due to psoriasis.

Patient Global Assessment (PtGA) of psoriasis

The five-point, patient-reported PtGA was a single-item question (Fig. 1b) that used the same category labels as the PGA (but without morphologic descriptors) and asked the patient to reflect their overall impression of their disease severity at the current time. The PtGA was completed at Baseline, Weeks 4, 8, 12, and 16, and early termination [3, 15].

PASI

PASI quantitates the severity of a patient’s psoriasis based on both lesion severity and the percent of body surface area (BSA) affected. PASI is a composite scoring by the investigator of degree of erythema, induration, and scaling (each scored separately) for each of four body regions, with adjustment for the percent of BSA involved for each body region and for the proportion of the body region to the whole body [1]. PASI scores range from 0 to 72, and higher PASI scores represent increasing psoriasis severity. Assessments were scheduled at the same times as the PGA and also at screening.

Pain/discomfort assessment (PDA)

Pain or discomfort in the joints or back (or both) was assessed at Baseline, Weeks 4, 8, and 12, and at early termination with the patient-reported PDA, a 0 (‘No pain or discomfort’) to 10 (‘Worst possible pain or discomfort’) numeric rating scale.

Ocular Comfort Index (OCI)

The OCI [16] investigated if symptoms of dry eye were increased in psoriasis and if treatment with tofacitinib affected these symptoms. This 12-item questionnaire was available in English and completed at Baseline, Week 12,

and at early termination. Patients assessed the frequency and severity of the six most common dry eye symptoms over the last week, with each item measured on a seven-point Likert scale. The OCI was scored using a polytomous Rasch model, according to the developer’s instructions. The OCI score ranges from 0 to 100, with a higher score indicating more severe ocular discomfort. A change of 3-points or greater is considered to be clinically meaningful [17].

Psychometric analyses

Confirmatory factor analysis with constrained path coefficients

Confirmatory factor analysis (CFA) tested the fit of the PGA measurement model. To model the scoring algorithm, which assumed equal weight for every item to create an observed PGA score, we constrained all paths from the latent PGA construct to the observed items (the ‘loadings’) to be equal to each other (Fig. 2a). If the effect of the latent or underlying PGA construct on each item is equal, then each item is influenced equally by the PGA construct, and each item provides equal weight in creating a PGA score that estimates the PGA construct. Bentler’s Comparative Fit Index (CFI) [18] was used as a measure of fit. Acceptable fit was defined as follows: (a) CFI > 0.9; (b) statistically significant path coefficients; and (c) standardized path coefficients greater than or equal to 0.4 [19–21].

Confirmatory factor analysis without constrained path coefficients

Constraints were relaxed by allowing loadings to be different in order to test statistical equality of the loadings (Fig. 2b). This constituted a ‘just-identified model’ because the number of parameters to be estimated was exactly equal to the number of known data points (based on the number of variances and covariances in the model). Therefore, there was no test for goodness of fit, as a ‘just-identified model’ has perfect fit. A total of 10,000 bootstrap simulations were performed (at every time point) to obtain 95 % confidence intervals (CIs). To assess statistical inferences, we examined 95 % CIs for differences between loadings [19–21].

Test–retest reliability

Instrument stability was assessed by administering the instrument to patients on at least two different occasions under a stable set of conditions. To assess PGA test–retest reliability, an intraclass correlation coefficient (ICC) was estimated based on the PGA data at Baseline and Week 2. A stable subgroup of patients was defined using PASI

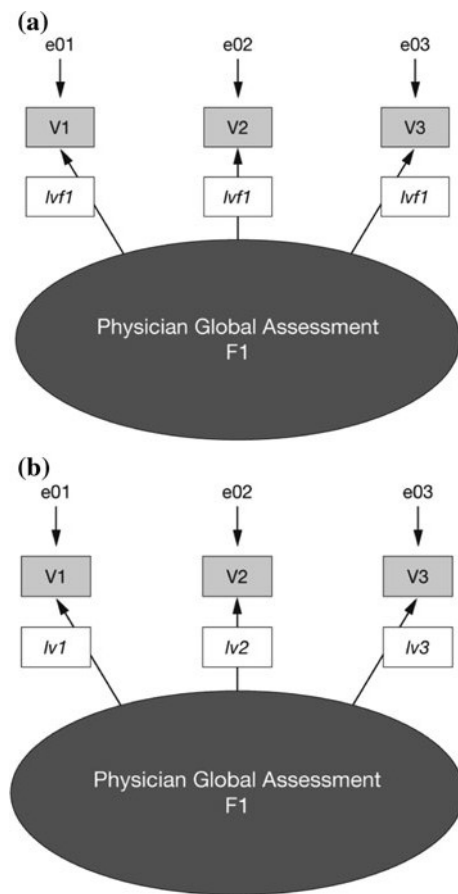


Fig. 2 Model of the confirmatory factor analysis: **(a)** with path coefficients constrained to be equal (representing current PGA scoring algorithm), **(b)** without constraints. The variables V1, V2, and V3 represent the three observed variables (items)—erythema, induration, and scaling—intended to measure the PGA (the latent variable F1). The variables e01, e02, and e03 represent the measurement errors associated with the variables V1, V2, and V3, respectively. The variable *lvf1* represents the path coefficient from the latent PGA construct (variable F1) to the item 1 (variable V1). In model **(a)**, the loadings are constrained to be equal for all three paths and are represented by the same variable *lvf1*. In model **(b)**, loadings are free to vary, with *lv1*, *lv2*, and *lv3* serving as path coefficients from the latent PGA construct to the items 1, 2, and 3, respectively. *PGA* Physician Global Assessment

scores. Patients with unchanged PASI scores between Baseline and Week 2 were used to estimate the ICC. Such a constraint of no change was relaxed if necessary to obtain a large enough sample for the ICC model to converge. An ICC > 0.70 was considered satisfactory [22].

Internal consistency reliability

Cronbach's coefficient α (CC α) and corrected item-to-total correlations (correlation of an item with the total score excluding that item) at every time point from Baseline to Week 16 were used to evaluate the extent to which the individual items of the PGA instrument were consistent

with each other. A CC α > 0.7 was considered acceptable [22].

Clinically important difference

Clinically Important Difference (CID) for the PGA was defined by employing a repeated measures model to estimate the relationship between PGA and PtGA score. While a meaningful and suitable anchor measure can come from different sources (e.g., clinical measure, clinician-rated, patient-rated), use of PtGA as an anchor is desirable because it captures directly the patient's experience on overall cutaneous disease. A linear relationship was imposed where PGA was the outcome and PtGA was a continuous predictor, serving as an anchor. This model incorporated all available data from Baseline and Weeks 4, 8, and 12.

The estimated CID was calculated as the mean difference on the PGA for a one-category difference on the PtGA [with discrete categories 'severe', 'moderate', 'mild', 'almost clear', 'clear (no psoriasis)']. Given that each category on the PtGA offers a clinically distinctive and meaningful response, which is inherently imbued by the names of the discrete categories on severity assessment, responses to different categories on the PtGA also warrant distinction and, by implication, provide a clinically important difference. The smallest clinically measurable difference is between a pair of adjacent categories taken to be clinically different by definition (although a smaller difference may also be clinically important but the scale, with its restricted number of categories, is not equipped to detect it). The mean difference in scores on the targeted PRO between adjacent categories on an anchor measure is a natural way to determine a CID (though not necessarily the minimal CID) on a PRO and has been recommended and applied in other applications [23–26].

A sensitivity analysis was also performed using PtGA as a categorical anchor, which does not impose any functional relationship between an outcome and an anchor.

The standardized effect size (ES) was calculated as the same difference in mean changes divided by standard deviation of the PGA at baseline. A more negative effect size indicated a larger benefit from tofacitinib relative to placebo. The following interpretation has been given to the magnitude of an ES in absolute value: 'trivial' ($|ESI| < 0.20$), 'small' ($0.2 \leq |ESI| < 0.50$), 'moderate' ($0.5 \leq |ESI| < 0.80$), and 'large' ($|ESI| \geq 0.80$) [27].

Known-group validity

Known-group validity was assessed by employing a repeated measures model to assess the relationship between PGA and PASI. PGA score was used as the outcome and PASI score as

a categorical anchor (therefore no functional relationship was imposed). Seven categories were created—category 0 (observations when patient PASI score was exactly 0) to category 6 (observations when patient PASI score was in the range from 40 to 72). This model utilized all available data from Baseline and Weeks 2, 4, 8, and 12.

Convergent and divergent validity

Convergent and divergent validity were assessed by determining the correlation of the PGA with the PtGA, PASI, PDA, and OCI. Evidence for convergent validity was based on a Pearson correlation of 0.40 or higher, consistent with a meaningful correlation [28]. Evidence for divergent validity was based on a Pearson correlation <0.30, consistent with a less-than-medium association [27]. Correlations between 0.30 and 0.40 were taken as insufficient evidence to dismiss either convergent or divergent validity. It was expected that the PGA would show convergent validity with the PtGA and PASI, and divergent validity with the PDA and OCI.

Implementation

All available data were used in all analyses. All statistical analyses were performed using SAS[®] version 9.2 (SAS[®] Institute Inc. 2008).

Results

A total of 197 patients with moderate-to-severe chronic plaque psoriasis were randomized to tofacitinib 2 mg BID ($n = 49$), 5 mg BID ($n = 49$), 15 mg BID ($n = 49$), or placebo ($n = 50$). PGA response rates were significantly higher for all tofacitinib groups at Week 12 compared with placebo [14].

Confirmatory factor analysis with constrained path coefficients

Bentler's CFI was more than 0.9, path coefficients were statistically significant, and standardized path coefficients were more than 0.4 for all time points during the study (Table 1), indicating that the specified model fitted the data.

Confirmatory factor analysis without constrained path coefficients

All standardized loadings at all time points exceeded 0.4, with the exception of the path to item 1 (erythema) at baseline, which had a standardized loading of almost 0.4 (0.39). There were no significant differences between path

coefficients in the vast majority of the 21 comparisons (Table 2). There were two exceptions (bold in table): At Week 2, the path to erythema and the path to induration were statistically different; and at Week 14, the path to erythema and path to scaling were statistically different.

Test-retest reliability

Only 26 patients had Baseline and Week 2 PASI scores exactly equal to each other. Because of the small number of patients, the model did not converge; as a result, the ICC could not be estimated. Constraints on the difference between Baseline and Week 2 PASI scores were then relaxed from no change to changes <1 point, which provided a subgroup of 50 patients resulting in an ICC for PGA of 0.8. When constraints on change in PASI score from Baseline to Week 2 were relaxed further to <2 and <3 points, the ICCs were still at acceptable levels (0.75 and 0.71, respectively). Constraints relaxed to <4 point differences supplied an ICC of 0.67.

Internal consistency reliability

At all time points from Week 2 onwards, the PGA scale had a CC α value greater than 0.8 (Table 3). Corrected item-to-total correlations were high and consistent over time, with correlations generally larger than 0.7 after baseline.

CID

The CID for the PGA score (linked to one-category difference on the PtGA) was estimated to be approximately 0.52 [95 % Confidence Interval (CI): 0.47, 0.56]. Based on this result, it is recommended that the value of approximately 0.5 (half of a PGA category) be used as the PGA CID. Results from the sensitivity analysis (PtGA as a categorical anchor) produced an extremely close functional relationship to the results using PtGA as a continuous

Table 1 Physician Global Assessment (PGA) confirmatory factor analysis with constrained path coefficients

Time point	Observations (N)	CFI ^a	Standardized path coefficient
Baseline	197	0.91	0.56
Week 2	191	0.97	0.73
Week 4	189	1	0.79
Week 8	169	1	0.89
Week 12	154	1	0.83
Week 14	145	1	0.88
Week 16	144	1	0.88

^a Bentler's comparative fit index (CFI)

Table 2 Physician Global Assessment (PGA) confirmatory factor analysis without constrained path coefficients: Differences between path coefficients loadings

Time point	Difference between path coefficients		
	Erythema versus Induration (95 % CI)	Erythema versus Scaling (95 % CI)	Induration versus Scaling (95 % CI)
Baseline	−0.180 (−0.504, 0.006)	−0.158 (−0.356, 0.007)	0.022 (−0.233, 0.383)
Week 2	−0.177 (−0.323, −0.039)	−0.123 (−0.255, 0.004)	0.054 (−0.095, 0.202)
Week 4	−0.065 (−0.186, 0.043)	−0.009 (−0.121, 0.105)	0.056 (−0.062, 0.186)
Week 8	−0.012 (−0.117, 0.087)	−0.014 (−0.103, 0.079)	−0.001 (−0.099, 0.102)
Week 12	−0.011 (−0.124, 0.095)	−0.070 (−0.173, 0.034)	−0.058 (−0.155, 0.039)
Week 14	−0.012 (−0.112, 0.085)	−0.096 (−0.193, −0.004)	−0.084 (−0.178, 0.010)
Week 16	0.002 (−0.110, 0.108)	−0.055 (−0.164, 0.048)	−0.057 (−0.170, 0.055)

Bold text identifies significant differences between path coefficients for 2 out of the 21 comparisons; *CI* confidence interval

Table 3 Physician Global Assessment (PGA) internal consistency reliability: Cronbach's coefficient α and corrected item-to-total correlations

Time point	Cronbach's coefficient α	Corrected item-to-total correlation ^a		
		Erythema	Induration	Scaling
Baseline	0.59	0.31	0.47	0.43
Week 2	0.81	0.58	0.70	0.70
Week 4	0.87	0.72	0.79	0.77
Week 8	0.92	0.83	0.84	0.83
Week 12	0.91	0.78	0.85	0.84
Week 14	0.91	0.81	0.83	0.81
Week 16	0.91	0.82	0.82	0.84

^a Corrected item-to-total correlations show correlation of an item with corrected total score of the scale (corrected total score is calculated as the sum of all other items in the scale, excluding the item of interest)

anchor, thus supporting a linearity assumption for the relationship between PGA and PtGA (Fig. 3). Results of the longitudinal modeling for PGA with added interpretations of the differences in terms of ES are detailed in Table 4. Tofacitinib 15, 5, and 2 mg BID have a large effect (relative to placebo) from Week 2, Week 4, and Week 8 onwards, respectively.

Known-groups validity

A robust monotonic functional relationship between PASI and PGA was observed (Fig. 4). All PGA score differences corresponding to a PASI score of 0 (patients without active disease) and PASI score greater than 0 (patients with active disease) were statistically different and clinically relevant

(differences in absolute value were greater than the CID value of 0.52) (Table 5). These results indicate evidence of known-group validity.

Convergent and divergent validity

Convergent validity on the PGA was evidenced by its relatively high correlation with the PASI. All correlations between PGA and PASI scores, and PGA and PtGA scores, were more than 0.5, except for the baseline observation (Table 6). The PDA and OCI scales provided evidence of divergent validity on the PGA. All correlations between PGA and PDA, and between PGA and OCI, were <0.2 (Table 6).

Discussion

Following the release of the European Medicines Agency Committee for Medicinal Products for Human Use (CHMP) guidance on psoriasis issued in 2004 [3], the clinical trial community did indeed follow the strong recommendation to include both PGA and PASI endpoints to assess treatment efficacy. In spite of the popularity of the PGA as a primary/co-primary/secondary efficacy endpoint for psoriasis clinical trials, the validity and reliability of the PGA have remained contentious [5–7]. This is in part due to the lack of a standard version of the PGA: PGA scales vary in the number of categories, the descriptors for each category, and whether it is a static or dynamic scale. In this paper, we used Phase 2 study data to provide a robust validation of a current version of this tool.

CFA tested the appropriateness of the PGA scoring algorithm with equal weighting of the three items. CFA

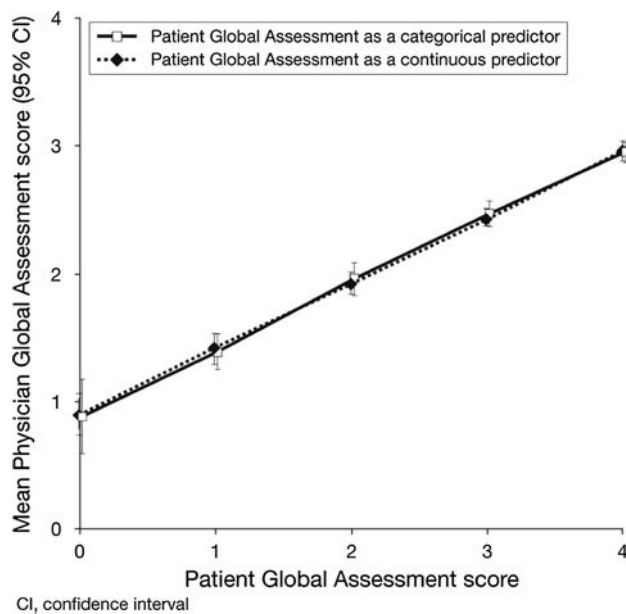


Fig. 3 Relationship between Physician Global Assessment score and Patient Global Assessment category. *CI* confidence interval

with path coefficients constrained to be the same value (representing the current PGA scoring algorithm) demonstrated excellent fit and provided statistically significant standardized path coefficients for all time points. CFA without constraints on path coefficients found no significant differences between path coefficients for a majority of comparisons. Thus, the modeling confirmed and validated

the current PGA scoring algorithm with equal weighting for the erythema, induration, and scaling items.

In evaluating the fit of the CFA model with constrained coefficients, we prospectively defined a good or acceptable fit as having Bentler's CFI > 0.9, as well as statistically significant path coefficients and standardized path coefficients greater than or equal to 0.4. Bentler's CFI was chosen because, of the several fit indexes available, it is regarded as the best one and the most popular and recommended index owing to its most favorable statistical properties [18–20]. Another recommended and well-behaved index is the non-normed fit index. When examined retrospectively, its values were essentially the same as those of the CFI.

In assessing test–retest reliability, it is recommended that ICC values exceed 0.70 [22]. Using a subgroup of patients who demonstrated exactly zero change in PASI score from Baseline to Week 2, the PGA ICC was not estimated due to the limited number of patients available for the modeling. Constraints were then relaxed in that the ‘no change in PASI score’ group was defined using increasing allowances of change in score (<1, <2, <3, and <4 points). Not until the group was permitted up to 4 points of change in PASI score did the PGA ICC fall below the acceptable level of 0.7. This evidence supports the test–retest reliability of the PGA.

Starting from Week 2, the PGA scale had excellent internal consistency reliability ($CC\alpha > 0.7$ in all but one calculation), and corrected item-to-total correlations were

Table 4 Physician Global Assessment (PGA) differences from placebo and corresponding effect sizes

Week	Tofacitinib dose (mg BID)	Difference from placebo ^a	Percentage of CID	Effect size ^b (classification)
Week 2	2	−0.24	48.0	−0.48 (small)
	5	−0.22	44.0	−0.44 (small)
	15	−0.44*	88.0	−0.89 (large)
Week 4	2	−0.25	50.0	−0.50 (moderate)
	5	−0.46*	92.0	−0.93 (large)
	15	−0.73**	146.0	−1.47 (large)
Week 8	2	−0.43*	86.0	−0.87 (large)
	5	−0.54*	108.0	−1.09 (large)
	15	−0.9**	180.0	−1.81 (large)
Week 12	2	−0.55*	110.0	−1.11 (large)
	5	−0.58*	116.0	−1.17 (large)
	15	−1.26**	252.0	−2.54 (large)

BID, twice daily; CID clinically important difference, SD standard deviation

* $p < 0.05$; ** $p < 0.0001$, full analysis set

^a Difference in mean change from baseline between tofacitinib dose and placebo

^b Effect size (ES) was calculated as a ‘Difference from placebo’ divided by PGA SD at baseline (value of 0.50), where a more negative effect size indicated a larger benefit from tofacitinib relative to placebo; the following interpretation has been given to the magnitude of ES in absolute value: ‘trivial’ ($|ESI| < 0.20$), ‘small’ ($0.2 \leq |ESI| < 0.50$), ‘moderate’ ($0.5 \leq |ESI| < 0.80$), and ‘large’ ($|ESI| \geq 0.80$) [27]

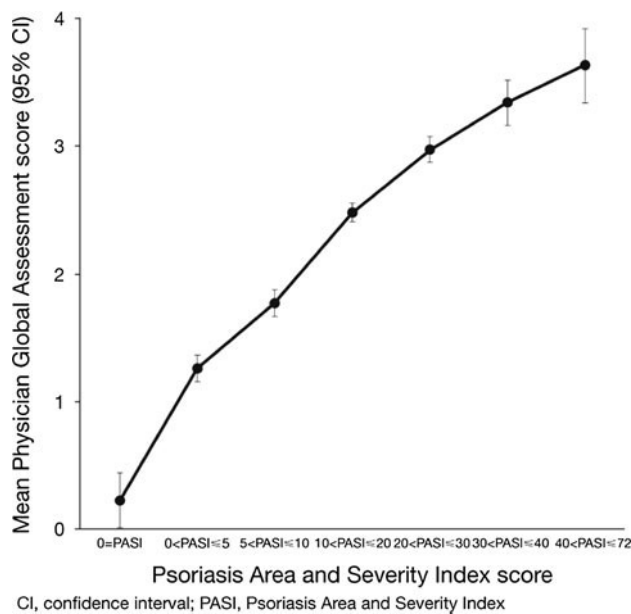


Fig. 4 Relationship between PASI score and Physician Global Assessment score. *CI* confidence interval, *PASI* Psoriasis Area and Severity Index

Table 5 Known group validity of PGA: Mean differences in PGA scores (from Baseline and Weeks 2, 4, 8, and 12) between the observations where PASI = 0 (Category 0) and observations where PASI score > 0 as indicated

Category definition based on the PASI score	Mean difference in PGA scores between group with PASI score = 0 and indicated PASI score group	95 % CI	
		Lower	Higher
(Category 0) PASI = 0			
(Category 1) 0 < PASI ≤ 5	−1.03	−1.25	−2.98
(Category 2) 5 < PASI ≤ 10	−1.54	−0.81	−2.52
(Category 3) 10 < PASI ≤ 20	−2.25	−1.77	−3.38
(Category 4) 20 < PASI ≤ 30	−2.75	−1.32	−2.84
(Category 5) 30 < PASI ≤ 40	−3.11	−2.47	−3.76
(Category 6) 40 < PASI ≤ 72	−3.40	−2.04	−3.05

All *p* values were <0.0001

PGA score was outcome, and PASI score was categorical anchor, with no functional relationship was imposed

CI confidence interval, *PASI* Psoriasis Area and Severity Index, *PGA* Physician Global Assessment

high and consistent over time. Relatively small $CC\alpha$ and corrected item-to-total correlations at baseline are most likely the result of the small dispersion of the data at

baseline due to the homogeneous pre-selection of the patients for the study, resulting in a restrictive range of scores. Because of the effect of treatment, there is a more diverse range of responses post-baseline, leading to higher correlations.

The correlations between PGA scores and PASI score groups at all time points except baseline can be considered as evidence for convergent validity of the PGA. At baseline, with a correlation of 0.31, there is insufficient evidence to dismiss either convergent validity or divergent validity. All post-baseline correlations between PGA and PtGA score groups were more than 0.5, which also supported convergent validity of the PGA. All correlations between PGA and PDA, and also between PGA and OCI, were <0.2 and can be considered as evidence of divergent validity on the PGA. Therefore, the PGA is correlating in ways that are expected—showing a high correlation with measures that are similar to it and a low correlation with measures that are dissimilar.

Modeling demonstrated a robust monotonic functional relationship between PASI and PGA (Fig. 4). The modeling capitalized on all available data simultaneously, in one cohesive and integrated model, by adjusting for the dependencies (or correlations) of the measures over time on the same set of individuals and, in doing so, provided an overall relationship between PGA and PASI that captured a consistent relationship across visits at baseline and weeks 2, 4, 8, and 12. These results also indicate known-groups validity: All differences in mean PGA scores corresponding to a PASI score of 0 (representing patients without active disease) and PASI score greater than 0 (representing patients with active disease) were statistically different and clinically relevant (all those differences in absolute value were larger than the estimated CID value of 0.5).

The CID for the PGA Score was estimated at 0.52 (95 % CI 0.47, 0.56), and, allowing for rounding error, a value of 0.5 (half of a PGA category) is proposed to be used as the PGA CID. It should be noted that, in the assessment of an individual patient, it is not possible to measure anything less than a one-category difference on the PGA, due to the limitation and configuration of its scale (which generally applies to other rating scales as well). The CID, though, pertains to a group difference and not an individual difference, making it suitable to evaluate a group effect.

Relative to placebo, all tofacitinib doses had PGA score improvements that were close to or exceeded the CID. This indicates that patients on tofacitinib experienced both statistically significant and clinically meaningful improvements in disease severity. It should be noted that an estimated CID is not necessarily a minimum CID, which is a more challenging and controversial area to pursue due to natural sampling variation, the type of anchor used and the time of assessment [24, 29]. If an estimated treatment

Table 6 Correlations to assess convergent validity and divergent validity

Correlation	Pearson correlation coefficients Prob > r under H0: Rho = 0 Number of observations						
	Baseline	Week 2	Week 4	Week 8	Week 12	Week 14	Week 16
PGA	0.31	0.54	0.67	0.74	0.79	0.75	0.74
versus	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
PASI	197	192	188	169	154	145	144
PGA	0.15	–	0.56	0.65	0.69	–	0.62
versus	0.0358		<0.0001	<0.0001	<0.0001		<0.0001
PtGA	197		187	170	153		138
PGA	–0.05	–	0.08	0.22	0.17	–	–
versus	0.4564		0.2990	0.0045	0.0315		
PDA	196		185	170	154		
PGA	–0.21	–	–	–	0.02	–	–
versus	0.0042				0.8140		
OCI Score	183				140		

PASI Psoriasis Area and Severity Index, *PDA* Pain/Discomfort Assessment, *PGA* Physician Global Assessment, *PtGA* Patient Global Assessment, *OCI* Ocular Comfort Index

effect (difference between groups) is not considerably less than an estimated CID, this does not imply that a meaningful CID has not occurred.

One asset of this study is that the CID for PGA was estimated based on a patient-reported measure, specifically, a one-category change on the PtGA. Fundamental to the patient-centered measurement of a CID on a target measure of interest, be it a subjective measure (subjective-based PGA from the clinician) or objective measure (e.g., blood pressure), is that its anchor should come from the patient when possible, which was followed in this study.

The PtGA asked the subject to evaluate the overall cutaneous disease at that point in time on a single-item five-point scale (severe, moderate, mild, almost clear, and clear). The PGA has the same set of response categories, assessed at the same point in time, as the PtGA. A patient-centered measure of important improvement is fundamental to the concept of CID when CID is quantified longitudinally, as it is here. While the PtGA is not a measure of patient-rated change per se (it a measure of a patient-rated assessment at the time of assessment), the repeated measures model employed regressed PGA (as the outcome or dependent variable for which a CID was to be estimated) on PtGA (as the anchor predictor) longitudinally. As such, by having repeated measurements over time on both PtGA and PGA scores for each subject, the longitudinal model incorporated and aligned patient-rated changes on the PtGA with physician-rated changes on the PGA. By relating mean changes on the PGA between adjacent categories on the PtGA at each time and also across time, the

longitudinal model calculated a CID on PGA that captured all available information appropriately. In essence, this anchor-based approach to determine a CID integrates or blends cross-sectional differences and longitudinal changes.

In general, a multi-item scale, with each item intended to measure a separate aspect of the same underlying construct, is more likely than a single-item measure to capture the richness and detail of complex constructs, such as the PGA of psoriasis severity. A multi-item scale is also more likely to benefit from increased precision, as random measurement of errors of different items cancel each other out [13]. Results of the analyses indicated that the three-item PGA scale has strong psychometric properties with clear evidence of being both reliable and valid, and that equal weighting of the three items (erythema, induration, scaling) is appropriate. Relative to single-item PGAs, this three-item scale is expected to be a more informative, accurate, and sensitive assessment of psoriasis severity. Although the PASI is widely used and has an established scoring system, this three-item PGA scale may prove to be of benefit due to its relative simplicity of interpretation and ease of use compared to the more complex and abstract PASI. The extent of the superiority of this PGA over single-item PGAs is a topic for future research.

Acknowledgments Editorial support was provided by Jonny Miller, PhD, and Anne Marie Reid, PhD, of Complete Medical Communications and was funded by Pfizer Inc. The authors would like to thank the A3921047 investigators, staff and patients, and the A3921047 study team.

Conflict of interest The study was funded by Pfizer Inc. Dr Cappelleri, Mr Bushmakina, and Dr Mamolo are employees of Pfizer Inc. Ms Harness was a full-time employee of Pfizer Inc during the conduct and reporting of the study and now works at Novartis Pharma AG, Basel, Switzerland. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

References

- Fredriksson, T., & Pettersson, U. (1978). Severe psoriasis—Oral therapy with a new retinoid. *Dermatologica*, 157, 238–244.
- Feldman, S. R., & Krueger, G. G. (2005). Psoriasis assessment tools in clinical trials. *Annals of the Rheumatic Diseases*, 64(Suppl 2), ii65–ii68.
- European Medicines Agency. (2004). *Guideline on clinical investigation of medicinal products indicated for the treatment of psoriasis*.
- Ashcroft, D. M., Wan Po, A. L., Williams, H. C., & Griffiths, C. E. (1999). Clinical measures of disease severity and outcome in psoriasis: A critical appraisal of their quality. *British Journal of Dermatology*, 141, 185–191.
- Langley, R. G., & Ellis, C. N. (2004). Evaluating psoriasis with psoriasis area and severity index, psoriasis global assessment, and lattice system physician's global assessment. *Journal of the American Academy of Dermatology*, 51, 563–569.
- Berth-Jones, J., Grotzinger, K., Rainville, C., et al. (2006). A study examining inter- and intrarater reliability of three scales for measuring severity of psoriasis: Psoriasis area and severity index, physician's global assessment and lattice system physician's global assessment. *British Journal of Dermatology*, 155, 707–713.
- Spuls, P. I., Lecluse, L. L. A., Poulsen, M.-L., Bos, J. D., Stern, R. S., & Nijsten, T. (2010). How good are clinical severity and outcome measures for psoriasis? Quantitative evaluation in a systematic review. *Journal of Investigative Dermatology*, 130, 933–943.
- Naldi, L. (2010). Scoring and monitoring the severity of psoriasis. What is the preferred method? What is the ideal method? Is PASI passé? facts and controversies. *Clinics in Dermatology*, 28(1), 67–72.
- Naldi, L., Svensson, A., Diepgen, T., et al. (2003). Randomized clinical trials for psoriasis 1977–2000: The EDEN survey. *Journal of Investigative Dermatology*, 120, 738–741.
- Robinson, A., Kardos, M., & Kimball, A. B. (2012). Physician global assessment and psoriasis area and severity index: Why do both? A systematic analysis of randomized controlled trials of biologics agents for moderate to severe plaque psoriasis. *Journal of the American Academy of Dermatology*, 66, 369–375.
- Puzenat, E., Bronsard, V., Prey, S., et al. (2010). What are the best outcome measures for assessing plaque psoriasis severity? A systematic review of the literature. *Journal of the European Academy of Dermatology and Venereology*, 24(Suppl 2), 10–16.
- Stern, R. S. (2010). Poor metrics and lost opportunity. *Journal of the American Academy of Dermatology*, 63, 718–719.
- Sloan, J. A., Aaronson, N., Cappelleri, J. C., et al. (2002). Assessing the clinical significance of single items relative to summated scores. *Mayo Clinic Proceedings*, 77, 479–487.
- Papp, K., Menter, A., Strober, B., et al. (2012). Efficacy and safety of tofacitinib, an oral JAK inhibitor, in the treatment of psoriasis: A Phase 2b randomised placebo-controlled dose-ranging study. *British Journal of Dermatology*, 167, 668–677.
- Garduno, J., Bhosle, M. J., Balkrishnan, R., & Feldman, S. R. (2007). Measures used in specifying psoriasis lesion(s), global disease and quality of life: A systematic review. *Journal of Dermatological Treatment*, 18, 223–242.
- Karabulut, A. A., Yalvac, I. S., Vahaboglu, H., et al. (1999). Conjunctival impression cytology and tear-film changes in patients with psoriasis. *Cornea*, 18, 544–548.
- Johnson, M. E., & Murphy, P. J. (2007). Measurement of ocular surface irritation on a linear interval scale with the ocular comfort index. *Investigative Ophthalmology & Visual Science*, 48, 4451–4458.
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 238–246.
- Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: The Guilford Press.
- Hatcher, L. (1994). *A step-by-step approach to using the SAS® system for factor analysis and structural equation modeling*. Cary, NC: SAS Institute Inc.
- Kline, R. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: The Guilford Press.
- Fayers, F. M., & Machin, D. (2007). *Quality of life: The assessment, analysis, and interpretation of patient-reported outcomes* (2nd ed.). Chichester, England: Wiley.
- Cappelleri, J. C., & Bushmakina, A. G. (2013). Interpretation of patient-reported outcomes. *Statistical Methods in Medical Research* (in press).
- King, M. T. (2011). A point of minimal important difference (MID): A critique of terminology and methods. *Expert Review of Pharmacoeconomics Outcomes Research*, 11, 171–184.
- Mulhall, J. O., Goldstein, I., Bushmakina, A., et al. (2007). Validation of the erectile hardness score. *The Journal of Sexual Medicine*, 4, 1626–1634.
- Deyo, R. A., Inui, T. S., Leininger, J., et al. (1982). Physical and psychosocial function in rheumatoid arthritis: Clinical use of a self-administered health status instrument. *Archives of Internal Medicine*, 142, 879–882.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Stevens, J. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Copay, A. G., Subach, B. R., Glassman, S. D., et al. (2007). Understanding the minimum clinically important difference: A review of concepts and methods. *The Spine Journal*, 7, 541–546.