EUROPEAN FOOT & ANKLE SOCIETY
Advancing, developing and educating in foot and ankle surgery

EFAS Score – Multilingual Development and Validation of a Patient-Reported Outcome Measure (PROM) by the Score Committee of the European Foot and Ankle Society (EFAS)

Martinus Richter [a, 1], Per-Henrik Agren[b, 2], Jean-Luc Besse[b, 3], Maria Cöster[b, 4], Hakon Kofoed[b, 5], Nicola Maffulli[b, 6], Dieter Rosenbaum [c, 7], Martijn Steultjens[d, 8], Fernando Alvarez [e, 9], Andrzej Boszczyk[e, 10], Kris Buedts[e, 11], Marco Guelfi [e, 12], Henryk Liszka[e, 13], Jan-Willem Louwerens[e, 14], Jussi P. Repo[e, 15], Elena Samaila[e, 16], Michael Stephens [e, 17], Angelique G. H. Witteveen [e, 14]

[a-e] Score Committee European Foot and Ankle Society, c/o European Foot and Ankle Society (EFAS), Brussels, Belgium

[a] Head and core member

[b] Core member

[c] Technical advisor and core member

[d] Outcome measure development expert and core member

[e] National affiliate member

[1] Department for Foot and Ankle Surgery Rummelsberg and Nuremberg, Schwarzenbruck, Germany

[2] Stockholms Fotkirurgklinik, Sophiahemmet University, Stockholm, Sweden

[3] Université Lyon 1, IFSTTAR, LBMC UMR-T 9406 – Laboratoire de Biomécanique et Mécanique des Chocs, Bron Cedex, France and Hospices Civils de Lyon, Centre Hospitalier Lyon-Sud, Service de Chirurgie Orthopédique et Traumatologique, Pierre-Bénite Cedex, France.

[4] Department of Clinical Sciences and Orthopedics, Skåne University Hospital, Malmö, Sweden and Department of Foot and Ankle Surgery, Capio Movement, Halmstad, Sweden

[5] Private praxis, Charlottenlund, Denmark

[6] Queen Mary University of London, Barts and The London School of Medicine and Dentistry, London, UK

[7] Movement Analysis Lab, Institute for Experimental Musculoskeletal Medicine University Hospital Muenster, Muenster, Germany

[8] School of Health and Life Sciences, Glasgow Caledonian University, Glasgow, Scotland, UK

[9] Orthopaedic Surgery Department. Sant Rafael Hospital. Barcelona. Spain

[10] Department of Traumatology and Orthopaedics, Centre of Postgraduate Medical Education, Otwock, Poland

[11] Foot and Ankle Surgery Unit, Orthopaedic Department ZNA Middelheim, Antwerpen, Belgium

[12] Orthopaedic department, Montallegro Clinic, Genoa, Italy; Foot & Ankle Department, “Policlinico di Monza” Salus Clinic, Alessandria, Italy and Foot & Ankle Surgery Consultant professor, Orthopaedic & traumatology specialization school, “D’Annunzio University”, Chieti-Pescara, Italy

[13] Department of Orthopaedics and Rehabilitation, University Hospital in Krakow, Poland

[14] Foot and Ankle Reconstruction Unit, Sint Maartenskliniek, Nijmegen, The Netherlands

[15] Department of Orthopedics and Traumatology, Central Finland Health Care District, Jyväskylä, Finland

[16] Orthopedic and Traumatology, University of Verona, Verona, Italy

[17] Mater Private Hospital, Dublin, Ireland

Corresponding author:

Score Committee European Foot and Ankle Society (EFAS)

European Foot and Ankle Society (EFAS)

280, Boulevard du Souverain - 1160 Brussels – Belgium

Abstract

Background

A scientifically sound validated foot and ankle specific score validated ab initio in different languages is missing. The aim of a project of the European Foot and Ankle Society (EFAS) was to develop, validate, and publish a new score (the „EFAS Score") conceived and validated in different European languages.

Methods

The EFAS Score was developed and validated in three stages: 1) item (question) identification, 2) item reduction and scale exploration, 3) confirmatory analyses and responsiveness. The following score specifications were chosen: scale/subscale (Likert 0-4), questionnaire based, outcome measure, patient related outcome measurement. For stage 3, data were collected pre-operatively and post-operatively at a minimum follow-up of 3 months and mean follow-up of 6 months. Item reduction, scale exploration, confirmatory analyses and responsiveness were executed using analyses from classical test theory and item response theory.

Results

Stage 1 resulted in 31 general and 7 sports related questions. In Stage 2, a 6-item general EFAS Score was constructed using English, German, French and Swedish language data. In Stage 3, internal consistency of the scale was confirmed in seven languages: the original four languages, plus Dutch, Italian and Polish (Cronbach's Alpha >0.86 in all language versions). Responsiveness was good, with moderate to large effect sizes in all languages, and significant positive association between the EFAS Score and patient-reported improvement.

No sound EFAS Sports Score could be constructed.

Conclusions

The multi-language EFAS Score was successfully validated in the orthopaedic ankle and foot surgery patient population, including a wide variety of foot and ankle pathologies. All score versions are freely available at here.

Keywords

Score; Foot; Ankle; Validation; PROM

Introduction

A scientifically sound validated foot and ankle specific outcome measure for different European languages is still missing. Indeed, language-specific cross-cultural validation in other languages than English is largely absent [1,2]. Some outcome measures were validated for specific pathologies such as hallux valgus, ankle arthritis or flatfoot [3-6]. The European Foot and Ankle Society (EFAS) established in 2013 a Score Committee to develop, validate, and publish a new score, the „EFAS Score", which is not specific for single pathologies for different European languages. The principal aim of this project was to develop and validate the EFAS Score simultaneously for different European languages.

Methods

Previous scores were analysed, and different types of scores were taken into consideration [1-31]. The EFAS patient-reported outcome measure (PROM), the 'EFAS Score', was developed and validated in three stages: 1) item identification, 2) item reduction and scale exploration, 3) confirmatory analyses and responsiveness.

Type of score

We aimed to develop a questionnaire-based PROM, with one or more subscales depending on the results of the development process. After discussing different types of rating scales, a 5-point Likert scale (0-4) was chosen.

Questions – Item identification

In the first stage, potentially relevant items from existing questionnaires were identified [1,2,4,6-30]. These items were combined into one pool of items that were taken forward into stage 2 of the development process. Given the low relevance of items related to sports activities for some diagnostic groups, it was decided at this point to develop two separate scores: a general item score and a sports-specific score. Additionally, to ensure comparability of outcomes, it was decided to use 5-point Likert rating scales for all items in the patient data collections for stages 2 and 3 of the process, regardless of the original scoring method of the questionnaire from which the item had been mutated. In total, 31 general items and 7 sports-specific items were taken forward into the second phase of the project.

Item reduction and scale exploration

Through a process of forward and backward translation performed by bilingual translators, the original English pool of 38 items was translated into German, French and Swedish. These four language versions were then used for the Stage 2 data collection. Participants were recruited from orthopaedic foot and ankle surgery departments. Inclusion criteria for participants were clinical and imaging indications for foot and ankle surgery and age ≥ 18 years. No exclusion criteria were used other than an inability to complete a written questionnaire. Data collection was performed in France, Germany,

Sweden and Ireland. In addition to providing an answer to each item on a 5-point scale, all participants also rated the relevance of the item to their situation on a 5-point scale.

Following data collection, the following analytic steps were taken to reduce the item pool into one general PROM and one sports PROM. All steps were performed separately for each language version and separately for the general and sports-specific items unless stated otherwise.

1. Items with a ceiling effect (i.e. already at the highest possible level for a large proportion of patients), low perceived relevance and a high proportion of missing values were noted and shortlisted for exclusion in subsequent steps.

2. Using all items, a principal component analysis (PCA) was performed. A PCA identifies clusters (principal components) of closely related items through a matrix of inter-item correlations. Principal components were retained into the next step if the eigenvalue > 1 and if it was located left of the elbow of the screen plot. Then, items were excluded from further analysis if they showed no clear association with any of the retained principal components, or if they showed a high association (item load > 0.40) on more than one principal component. Cronbach's Alpha was computed for each of the identified principal components to explore their reliability. Any item showing a detrimental effect on scale reliability (i.e., Cronbach's Alpha would improve if the item was removed) was then excluded from further analysis. Finally, any item showing an item-scale correlation of r<0.60 was excluded. At the end of this step, the remaining items in their respective principal components would provide optimal scale reliability according to classic test theory.

3. An Item-response theory (IRT) analysis was performed for each of the identified scales (i.e., principal components) to further reduce the number of items and optimize scale unidimensional. These analyses were performed combining available data from all language versions. Items were first excluded if they showed reverse thresholds. It is expected that for any valid item, the probability of providing a certain response is closely linked to the underlying level of the trait that is measured. The order in which each response is the most likely response is a logical sequence. Two examples are provided to illustrate this in Figure 1a-b. Figure 1a shows an item with no reverse thresholds: for each of the five responses to the item, a level of the underlying trait (on the x-axis) can be identified at which that response is the most likely response (as signified by the probability level on the y-axis)and the order in which the five responses are most likely is logically progressing from response 1 to response 5. In contrast, figure 1b shows an item for which only the two most extreme responses are ever the most likely. This item in 1b would not pass the test for reverse thresholds and would therefore be excluded from the scale.

Then, a backward elimination process was used. Starting from the full remaining pool of items, at each step the worst-performing item was excluded. This was based on three parameters:

- *The slope parameter (a)*. This parameter represents how well an individual item can differentiate between patients with different "true" levels of the trait that is being measured. Higher values indicate better discrimination, with a=0.3 as the minimum acceptable value.

- *The item difficulty parameter (b).* This parameter indicates how commonly and severely a specific health issue (i.e. individual item) is flagged up by patients due to their foot/ankle problem. E.g., if 'pain during walking' is more likely to occur than 'pain while at rest' this would

be reflected in different 'b' parameters for those two items, showing that more patients are experiencing more severe pain during walking than at rest. A value between -3.0 and +3.0 was deemed acceptable.

- *The p-value of the Chi-Square statistic*. A significant p-value shows that the item does not fit the IRT model. Essentially, an IRT model expects items to fit into one hierarchy, from 'easy' to 'difficult' items. If a patient identifies severe health issues by selecting an affirmative response to a 'difficult' item, then logically 'easier' items should also show an affirmative response given that they ask after closely related but mild- to-moderate health issues. A significant p-value flags up that an item does not fit into this hierarchy and therefore does not fit the scale.

The worst performing item was removed from further analyses. The item to be excluded at each step was identified through the following hierarchy of criteria:

1. a slope parameter of $a<0.3$.
2. an item difficulty parameter outside the accepted range of values, i.e. $b< -3.0$ or $b>3.0$.
3. If criteria a) and b) did not apply to any item, the item with the most significant p value of the Chi-Square statistic was removed. For this criterion, the significance level was Bonferroni corrected per analysis for an overall critical value of $p<0.05$.

- The process was concluded when all remaining items showed good fit to the IRT model by having acceptable slope and item difficulty parameters and a non-significant p value for the Chi-square test.

Confirmatory analysis and responsiveness

Data collection for this final stage took place in the four original language versions, as well as Dutch, Italian and Polish.

Inclusion criteria for participants were scheduled foot and ankle surgery and age ≥ 18 years. No exclusion criteria were used other than an inability to complete a written questionnaire. Data was collected preoperatively and at postoperative follow-up. Minimum postoperative follow-up of 3 months and mean follow-up of 6 months was postulated (Table 1). As necessary score sheet numbers, 500 were postulated for two languages and 100 for all others (Table 1). To confirm the internal consistency for each language version, Cronbach's Alpha of the EFAS Score was computed for each language version separately. To establish the responsiveness of the EFAS Scores, both distribution-based and criterion-based analyses were used. Distribution-based measures of responsiveness included the effect size (ES) and minimal important difference (MID). The criterion-based measure of responsiveness used was the linear association (Pearson's correlation) between improvement on the EFAS Score and a 5-point Likert scale anchor question: did the surgery improve the foot and/or ankle problem? (0= no, not at all; 4 = Yes, very much).

The ES was calculated as the difference between the baseline and 6-month follow-up mean EFAS Score, divided by the standard deviation of the baseline EFAS Score:

The MID was considered to be equal to the standard error of measurement (SEM) of the baseline EFAS Score. The SEM was calculated as:

(Formula 1), where:

SD = standard deviation of the EFAS Score baseline score

r = value of Cronbach's Alpha for the EFAS Score at baseline.

To assess the responsiveness of the EFAS Score using the MID, the percentage of participants with an improvement in their EFAS Score between baseline and follow-up exceeding the MID was identified.

Statistical analyses were performed in SPSS (IBM SPSS Statistics 23, IBM, Armonk, NY, USA). The IRT modelling was performed in XCalibre 4 (Assessment Systems, Inc.)

Ethics

Approvals from the relevant ethical committees in different contributing countries were obtained, adhering to all local legislation.

Results

Table 1 and Table 2 and show the language-specific demographic data (Table 1) and diagnoses (Table 2) for the patient samples in stages 2 and 3 of the development process.

Item reduction

The 31 general items showed a wide spread of perceived relevance, proportion of missing values and floor and ceiling effects. Full descriptive for the four languages can be found in the supplementary materials (Tables S1a-d). Then, the PCAs were performed and item contribution to the identified subscales was analysed. Full results of the PCA for each of the four languages (scree plots, item loading on the principal components, item-scale correlations and Cronbach's alpha if item deleted) can be found in the supplementary materials, Figures S1a-d, Tables S2a-d and S3a-d. Table 3 summarises the findings, indicating which items were retained for further analyses and which items were dropped due to any of the previously stated criteria.

In two languages, French and English, the dominant principal component included both pain and physical function items. In the two other languages (German and Swedish), pain and physical function were identified as two separate principal components. In both cases, however, the correlation between those two subscales was significant: r=0.66 in German, and r=0.28 in Swedish.

Based on these findings, the retained items were taken forward into the IRT modelling with the aim to construct a single scale including both pain and physical function items. Additionally, to preserve the potential of the final EFAS Score to cover other issues relevant to patients than pain and physical function, two items relating to footwear were retained for the final step of the item reduction process. In total, 16 of the general items proceeded into the IRT modelling.

The results of the item reduction through the IRT analysis are provided in Table 4 for the general items.

This process resulted in a 6-item scale. Table 5 lists the six items in order of difficulty, providing insight into the hierarchy of items within the scale. The table shows that patients will most easily provide a low score to Q17 (indicating severe issues with pain during physical activity) while pain at rest (Q1) is least likely to be flagged up as a problem by patients. Both footwear items, retained after the previous step, were excluded in the first step of IRT modelling due low relevance.

Confirmatory analyses and responsiveness

Using news samples of patients, and including three new languages (Dutch, Italian and Polish), the 6-item EFAS Score was then evaluated for internal consistency and responsiveness.

The internal consistency of the scale was excellent in all seven language versions. Cronbach's Alpha was 0.86 in German; 0.86 in French; 0.92 in English; 0.87 in Swedish; 0.91 in Italian; 0.88 in Dutch and 0.86 in Polish.

Responsiveness of the EFAS Score is shown in Table 6 and Figures 2a-g. Moderate (ES>0.5) or large (ES>0.8) effect sizes were found in all language versions. A clear majority of patients showed a minimally important difference following surgery, ranging from 57.6% in German to 94.1% in English. The change in EFAS Scores between baseline and follow-up was significantly correlated with the patient-reported change in health status. Figures 3a-g show this positive association.

EFAS Sports Score

The separate EFAS Sports Score was developed similarly to the general EFAS Score. For the sports-related items (QS1-QS7), descriptives from stage 2 of the process can be found in supplementary Tables S1a-d. The PCA found one principal component for these items in all four languages. Based on the descriptives and the outcome of the PCA, all seven sports-related items were retained into the IRT modelling step, with an aim to construct a single EFAS Sports Score.

Table 7 shows the results of the IRT modelling on the sports-related items. No scale meeting the requirements of the IRT model could be constructed in these analyses. Due to this result, no confirmatory and responsiveness analyses were performed for the EFAS Sports Score.

Discussion

A six-item, single-scale EFAS Score was successfully developed and validated simultaneously in seven European languages. This outcome measure covers pain and physical function and was found to be internally consistent, unidimensional and responsive to change in samples of orthopaedic foot and ankle surgery patients. The maximum score is 24 points (best possible), and the minimum 0 points (worst possible).

Many foot and ankle scores were developed previously [1,2,4,6-30]. However, all of the above cited scores were developed and/or validated in English except VAS FA (Visual Foot and Ankle Scale, English, German, Italian, Thai) and SEFAS (Self-Reported Foot and Ankle Score, Swedish version) [1,2,32]. The language-specific validation was necessary because simple translation of a validated score does not necessarily result in a validated translated score [1,2]. This issue is especially important

for Europe with numerous languages. The mostly spoken mother tongues in Europe are German (18%), English (13%), Italian (13%), French (12%), Spanish (8%), Polish (8%) and Dutch (4%) (source Wikipedia, February 20, 2018). Therefore, a need for different language-specific (validated) scores, especially in Europe, is clear. Another issue is that some scores are only validated for specific pathologies such as hallux valgus, ankle arthritis or flatfoot [3-6]. EFAS recognized the need and importance for a non-pathology specific score in different European languages and therefore undertook the process described in the present study. The core committee comprised foot and ankle surgeons from different European countries completed with technical advisor (PhD for biomechanics) and an expert in outcome measure development and validation [33-39]. Patient involvement in the development of the EFAS Score was twofold: items were selected from questionnaires that had been developed using patient statements, and in stage 2 of the process, all questions that were deemed to be of low relevance by patients were excluded from further consideration.

The process of item reduction started with a pool of 31 candidate items and finished with the definitive 6-item EFAS Score. Using an analytical approach from classical test theory, the pool was reduced to 15 items while identifying the main underlying constructs of pain and physical function. Then, using item-response modelling, an unidimensional scale was identified comprising six items covering both pain and physical function. This scale had superior measurement properties over other variants that captured wider issues such as footwear and assistive device use, and was therefore preferred. Due to the simultaneous development of the EFAS Score in multiple languages, the development and validation process was highly complex. The analytical process was designed to do justice to the original aim of developing an outcome measure that is easy to use in clinical practice (i.e., short; relevant; generically applicable; valid and responsive) across different languages, health care systems and diagnostic groups, while maintaining optimal scientific credibility.

Not all measurement properties of the EFAS Score have been established. In particular test-retest reliability, i.e. reproducibility of the score in a stable (pre-surgery) population, was not included in the present study. As a result, the MID of the EFAS Score as reported in this study was based on the internal consistency of the scale (Cronbach's Alpha) rather than test-retest reliability. In future, if the test-retest reliability becomes available, this may lead to an adjustment in the SEM and therefore MID of the EFAS Score.

The process to develop the EFAS Sports Score was ultimately unsuccessful. The questions related to sports activities were not relevant to a large proportion of the patient samples, and suffered from a high proportion of missing values. This meant the IRT modelling did not result in a unidimensional EFAS Sports Score. Based on the findings of the IRT model, a 4-item EFAS Sports Score could be considered, as this was the best-performing option. In that scale, three items fit the scale well with a fourth item showing some issues. Removing the fourth item, however, then resulted in the remaining three items not forming a scale that fit the constraints of the IRT model. Further analyses in specific patient groups, for whom sports-related questions are relevant, is warranted.

In conclusion, the EFAS Score has been successfully validated for orthopaedic foot and ankle surgery populations incorporating a wide variety of foot and ankle pathologies, including language-specific validation in seven languages so far (English, German, French, Swedish, Dutch, Italian, Polish). Validation for other languages (Finnish, Portuguese, Spanish) is in progress. All validated score versions are freely available here.

References

1. Richter M, Zech S, Geerling J, Frink M, Knobloch K, Krettek C. A new foot and ankle outcome score: Questionnaire based, subjective, Visual-Analogue-Scale, validated and computerized. *Foot Ankle Surg* 2006*;* 12(4): 191-9.

2. Coster M, Karlsson MK, Nilsson JA, Carlsson A. Validity, reliability, and responsiveness of a self-reported foot and ankle score (SEFAS). *Acta Orthop* 2012*;* 83(2): 197-203.

3. Chen L, Lyman S, Do H, Karlsson J, Adam SP, Young E, Deland JT, Ellis SJ. Validation of foot and ankle outcome score for hallux valgus. *Foot Ankle Int* 2012*;* 33(12): 1145-55.

4. Roos EM, Brandsson S, Karlsson J. Validation of the foot and ankle outcome score for ankle ligament reconstruction. *Foot Ankle Int* 2001*;* 22(10): 788-94.

5. Mani SB, Brown HC, Nair P, Chen L, Do HT, Lyman S, Deland JT, Ellis SJ. Validation of the Foot and Ankle Outcome Score in adult acquired flatfoot deformity. *Foot Ankle Int* 2013*;* 34(8): 1140-6.

6. Madeley NJ, Wing KJ, Topliss C, Penner MJ, Glazebrook MA, Younger AS. Responsiveness and Validity of the SF-36, Ankle Osteoarthritis Scale, AOFAS Ankle Hindfoot Score, and Foot Function Index in End Stage Ankle Arthritis. *Foot Ankle Int* 2012*;*33(1): 57-63.

7. Kitaoka HB, Alexander IJ, Adelaar RS, Nunley JA, Myerson MS, Sanders M. Clinical rating systems for the ankle-hindfoot, midfoot, hallux, and lesser toes. *Foot Ankle Int* 1994*;* 15(7): 349-53.

8. Dawson J, Coffey J, Doll H, Lavis G, Cooke P, Herron M, Jenkinson C. A patient-based questionnaire to assess outcomes of foot surgery: validation in the context of surgery for hallux valgus. *Qual Life Res* 2006*;* 15(7): 1211-22.

9. Hung M, Nickisch F, Beals TC, Greene T, Clegg DO, Saltzman CL. New paradigm for patient-reported outcomes assessment in foot & ankle research: computerized adaptive testing. *Foot Ankle Int* 2012*;* 33(8): 621-6.

10. SooHoo NF, McDonald AP, Seiler JG, III, McGillivary GR. Evaluation of the construct validity of the DASH questionnaire by correlation to the SF-36. *J Hand Surg [Am]* 2002*;* 27(3): 537-41.

11. SooHoo NF, Shuler M, Fleming LL. Evaluation of the validity of the AOFAS Clinical Rating Systems by correlation to the SF-36. *Foot Ankle Int* 2003*;* 24(1): 50-5.

12. Budiman-Mak E, Conrad KJ, Roach KE. The Foot Function Index: a measure of foot pain and disability. *J Clin Epidemiol* 1991*;*44(6): 561-70.

13. Kuyvenhoven MM, Gorter KJ, Zuithoff P, Budiman-Mak E, Conrad KJ, Post MW. The foot function index with verbal rating scales (FFI-5pt): A clinimetric evaluation and comparison with the original FFI. *J Rheumatol* 2002*;* 29(5): 1023-8.

14. Saag KG, Saltzman CL, Brown CK, Budiman-Mak E. The Foot Function Index for measuring rheumatoid arthritis pain: evaluating side-to-side reliability. *Foot Ankle Int* 1996*;* 17(8): 506-10.

15. Domsic RT, Saltzman CL. Ankle osteoarthritis scale. *Foot Ankle Int* 1998*;* 19(7): 466-71.

16. Westphal T, Piatek S, Halm JP, Schubert S, Winckler S. Outcome of surgically treated intraarticular calcaneus fractures–SF-36 compared with AOFAS and MFS. *Acta Orthop Scand* 2004*;* 75(6): 750-5.

17. Grant S, Aitchison T, Henderson E, Christie J, Zare S, McMurray J, Dargie H. A comparison of the reproducibility and the sensitivity to change of visual analogue scales, Borg scales, and Likert scales in normal subjects during submaximal exercise. *Chest* 1999*;* 116(5): 1208-17.

18. Jamison RN, Gracely RH, Raymond SA, Levine JG, Marino B, Herrmann TJ, Daly M, Fram D, Katz NP. Comparative study of electronic vs. paper VAS ratings: a randomized, crossover trial using healthy volunteers. *Pain* 2002*;* 99(1-2): 341-7.

19. Ohnhaus EE, Adler R. Methodological problems in the measurement of pain: a comparison between the verbal rating scale and the visual analogue scale. *Pain* 1975*;* 1(4): 379-84.

20. SooHoo NF, Samimi DB, Vyas RM, Botzler T. Evaluation of the validity of the Foot Function Index in measuring outcomes in patients with foot and ankle disorders. *Foot Ankle Int* 2006*;* 27(1): 38-42.

21. Hale SA, Hertel J. Reliability and Sensitivity of the Foot and Ankle Disability Index in Subjects With Chronic Ankle Instability. *J Athl Train* 2005*;* 40(1): 35-40.

22. Hunt KJ, Hurwit D. Use of patient-reported outcome measures in foot and ankle research. *J Bone Joint Surg Am* 2013*;* 95(16): e118-e9.

23. Morssinkhof ML, Wang O, James L, van der Heide HJ, Winson IG. Development and validation of the Sports Athlete Foot and Ankle Score: an instrument for sports-related ankle injuries. *Foot Ankle Surg* 2013*;* 19(3): 162-7.

24. Coster MC, Rosengren BE, Bremander A, Brudin L, Karlsson MK. Comparison of the Self-reported Foot and Ankle Score (SEFAS) and the American Orthopedic Foot and Ankle Society Score (AOFAS). *Foot Ankle Int* 2014*;* 35(10): 1031-6.

25. Enneking WF, Dunham W, Gebhardt MC, Malawar M, Pritchard DJ. A system for the functional evaluation of reconstructive procedures after surgical treatment of tumors of the musculoskeletal system. *Clin Orthop Relat Res* 1993*;* 286): 241-6.

26. Button G, Pinney S. A meta-analysis of outcome rating scales in foot and ankle surgery: is there a valid, reliable, and responsive system? *Foot Ankle Int* 2004*;* 25(8): 521-5.

27. Agel J, Beskin JL, Brage M, Guyton GP, Kadel NJ, Saltzman CL, Sands AK, Sangeorzan BJ, SooHoo NF, Stroud CC, Thordarson DB. Reliability of the Foot Function Index:: A report of the AOFAS Outcomes Committee. *Foot Ankle Int* 2005*;* 26(11): 962-7.

28. Martin RL, Irrgang JJ, Burdett RG, Conti SF, Van Swearingen JM. Evidence of validity for the Foot and Ankle Ability Measure (FAAM). *Foot Ankle Int* 2005*;* 26(11): 968-83.

29. Ramanathan AK, Herd F, Macnicol M, Abboud RJ. A new scoring system for the evaluation of clubfoot: the IMAR-Clubfoot scale. *Foot (Edinb)* 2009*;* 19(3): 156-60.

30. Hung M, Baumhauer JF, Brodsky JW, Cheng C, Ellis SJ, Franklin JD, Hon SD, Ishikawa SN, Latt LD, Phisitkul P, Saltzman CL, SooHoo NF, Hunt KJ. Psychometric Comparison of the

PROMIS Physical Function CAT With the FAAM and FFI for Measuring Patient-Reported Outcomes. *Foot Ankle Int* 2014*;* 35(6): 592-9.

31. Hunt KJ, Alexander I, Baumhauer J, Brodsky J, Chiodo C, Daniels T, Davis WH, Deland J, Ellis S, Hung M, Ishikawa SN, Latt LD, Phisitkul P, SooHoo NF, Yang A, Saltzman CL. The Orthopaedic Foot and Ankle Outcomes Research (OFAR) network: feasibility of a multicenter network for patient outcomes assessment in foot and ankle. *Foot Ankle Int* 2014*;* 35(9): 847-54.

32. Stuber J, Zech S, Bay R, Qazzaz A, Richter M. Normative data of the Visual Analogue Scale Foot and Ankle (VAS FA) for pathological conditions. *Foot Ankle Surg* 2011*;* 17(3): 166-72.

33. Woodburn J, Vliet Vlieland TP, van der Leeden M, Steultjens MP. Rasch analysis of Dutch-translated version of the Foot Impact Scale for rheumatoid arthritis. *Rheumatology (Oxford, England)* 2011*;* 50(7): 1315-9.

34. Wiertsema SH, van Hooff HJ, Migchelsen LA, Steultjens MP. Reliability of the KT1000 arthrometer and the Lachman test in patients with an ACL rupture. *Knee* 2008*;* 15(2): 107-10.

35. Wiertsema SH, de Witte PB, Rietberg MB, Hekman KM, Schothorst M, Steultjens MP, Dekker J. Measurement properties of the Dutch version of the Western Ontario Shoulder Instability Index (WOSI). *J Orthop Sci* 2014*;* 19(2): 242-9.

36. Steultjens MP, Stolwijk-Swuste J, Roorda LD, Dallmeijer AJ, van Dijk GM, Post B, Dekker J. WOMAC-pf as a measure of physical function in patients with Parkinson's disease and late-onset sequels of poliomyelitis: unidimensionality and item behaviour. *Disability and rehabilitation* 2012*;* 34(17): 1423-30.

37. Peter WF, Steultjens MP, Mesman T, Dekker J, Hoeksma AF. Interobserver reliability of the Amsterdam Severity Scale in Stenosing Tenosynovitis (ASSiST). *Journal of hand therapy : official journal of the American Society of Hand Therapists* 2009;22(4): 355-9; quiz 60.

38. Peter WF, de Vet HCW, Boers M, Harlaar J, Roorda LD, Poolman RW, Scholtes VAB, Steultjens M, Hendry GJ, Roos EM, Guillemin F, Benedetti MG, Cavazzuti L, Escobar A, Dagfinrud H, Terwee CB. Cross-Cultural and Construct Validity of the Animated Activity Questionnaire. *Arthritis care & research* 2017*;* 69(9): 1349-59.

39. Eyssen IC, Steultjens MP, Oud TA, Bolt EM, Maasdam A, Dekker J. Responsiveness of the Canadian occupational performance measure. *Journal of rehabilitation research and development* 2011*;* 48(5): 517-28.

Legend

Figure 1a and 1b. Cumulative Response Functions for an item showing no reverse thresholds (1a) and an item showing reverse thresholds (1b). X-axis: 'true level of underlying trait; y-axis: probability of response.

Figure 2a-g: Association between change in EFAS Score from pre- to post-surgery and patient self-reported improvement (a, German; b, French; c, English; d, Swedish; e, Dutch; f, Italian; g, Polish)

Table 1. Demographic data. N = sample size; F = Female; L/R/B = Left/Right/Both; N/A = not available

Table 2. Prevalence of primary diagnoses, in %, based on ICD-10 codes

Table 3. Summary of results of the PCA. For the four individual language columns: √ = item retained, <blank> = item excluded. For the Overall column: Y = item retained in all four languages, ~ = item retained in three languages, <blank> = item retained in <3 languages.

Table 4. Results of the item reduction through the IRT analysis. RT = reverse thresholds.

Table 5. Items included in EFAS Score based on IRT modelling with their item difficulty parameter

Table 6. Responsiveness of the EFAS Score.

Table 7. Results of the IRT modelling on the sports-related items. K = number of items in analysis; RT = reverse thresholds; OK = item retained for next step of analysis; a significant P-value (P<0.05) indicates lack of fit to the IRT model.