J Rehabil Med 2018; 50: 569–574

ORIGINAL REPORT

# EVALUATION OF THE WEAR-AND-TEAR SCALE FOR THERAPEUTIC FOOTWEAR, RESULTS OF A GENERALIZABILITY STUDY

Rutger DAHMEN, MD[1], Petra C. SIEMONSMA, PT, PhD[1], Sandra MONTEIRO, PhD[2], Geoffrey R. NORMAN, PhD[2], Maarten BOERS, MD, PhD[3], Gustaaf J. LANKHORST, MD, PhD[4] and Leo D. ROORDA, MD, PhD, PT[1]

*From the [1]Amsterdam Rehabilitation Research Center | Reade, Amsterdam, The Netherlands, [2]Department of Health Research Methods Evidence and Impact, McMaster University, Hamilton, Ontario, Canada, [3]Department of Epidemiology and Biostatistics, Amsterdam Rheumatology and Immunology Center, and [4]Department of Rehabilitation, VU University Medical Center, Amsterdam, The Netherlands*

***Objective:*** **Therapeutic footwear is often prescribed at considerable cost. Foot-care specialists normally assess the wear-and-tear of therapeutic footwear in order to monitor the adequacy of the prescribed footwear and to gain an indicator of its use. We developed a simple, rapid, easily applicable indicator of wear-and-tear of therapeutic footwear: the wear-and-tear scale. The aim of this study was to investigate the intra- and inter-rater reliability of the wear-and-tear scale.**

***Methods:*** **A test set of 100 therapeutic shoes was assembled; 24 raters (6 inexperienced and 6 experienced physiatrists, and 6 inexperienced and 6 experienced orthopaedic shoe technicians) rated the degree of wear-and-tear of the shoes on the scale (range 0–100) twice on 1 day with a 4-h interval (short-term) and twice over a 4-week interval (long-term). Generalizability theory was applied for the analysis.**

***Results:*** **Short-term, long-term and overall intra-rater reliability was excellent (coefficients 0.99, 0.99 and 0.98; standard error of measurement (SEM) 2.6, 2.9 and 3.9; smallest detectable changes (SDC) 7.3, 8.0 and 10.8, respectively). Inter-rater reliability between professions, experience and inexperienced raters, and overall was excellent (coefficients 0.97, 0.98 and 0.93; SEM 4.9, 4.5, and 8.1; SDC 13.7, 12.4 and 22.5, respectively).**

***Conclusion:*** **The wear-and-tear scale has excellent intra-rater, inter-rater, and overall reliability.**

*Key words:* therapeutic footwear; use; wear-and-tear; generalizability theory; reliability.

Accepted Mar 7, 2018; Epub ahead of print May 16, 2018

J Rehabil Med 2018; 50: 569–574

Correspondence address: Rutger Dahmen, Amsterdam Rehabilitation Research Center | Reade, PO Box 58271, 1040 HG Amsterdam, The Netherlands. E-mail: r.dahmen@reade.nl

## MAIN MESSAGE

People with conditions such as diabetes mellitus, rheumatoid arthritis and osteoarthritis often experience foot problems that cause pain and abnormal form and functioning, leading to limitations in standing and walking. In the Netherlands, therapeutic footwear (TF) is part of the usual care prescribed and controlled at the multidisciplinary foot-care clinic by physiatrists and orthopaedic shoe technicians. These professionals judge the wear-and-tear of TF in order to monitor the adequacy of the prescribed footwear, to gain an indicator of its use, and to determine the necessity for a new pair of these expensive TF. The aim of this study was to evaluate a tool for measuring the wear-and-tear of TF: the wear-and-tear scale.

The estimated prevalence of foot problems in the general population varies from 26% to 30% (1). Foot problems are common features of chronic diseases such as diabetes mellitus (2), rheumatoid arthritis (RA) (3) and osteoarthritis (4). The prevalence of foot problems in RA varies from 50% to 90% (5, 6) and in osteoarthritis from 12% to 17%. Foot problems include foot pain, abnormal foot posture and function, which lead to limitations in weight-bearing activities, such as standing and walking (3, 7, 8). This, in turn, contributes to restrictions in participation in life situations (9–11). Therapeutic footwear (TF), as part of the usual non-pharmacological care for foot problems (12), is often prescribed at considerable cost. As an example from the Netherlands, this cost is approximately 60 million Euros per year (13).

In daily clinical practice, foot-care specialists and other healthcare professionals involved in the prescription of TF normally assess the wear-and-tear of TF. Firstly, they do so to monitor the adequacy of the prescribed footwear. If, for instance, there is a pattern of wear-and-tear at a specific position on the footwear, this may indicate that the prescription is suboptimal and should be improved for current or future footwear. Secondly, they assess wear-and-tear of TF in order to gain an indicator of its use and, related to this, to determine the necessity for a new pair of TF.

This raises the questions of how foot-care specialists and other professionals involved in the prescription of TF can quantify the assessment of wear-and-tear, and, related to this, what is the reliability of such assessment. We have developed a wear-and-tear scale: a simple, rapid, clinically applicable and inexpensive indicator of the wear-and-tear of TF. The aim of the current study is to evaluate the intra- and inter-rater reliability of the wear-and-tear scale (14).


doi: 10.2340/16501977-2339

## METHODS

The study was reviewed by the ethics committee of Slotervaart Hospital and Reade, who undertook an expedited review and determined that the research activities described met their requirements of exemption from review.

### *Shoes*

A test set was assembled from 100 consecutive eligible therapeutic shoes (50 pairs) collected from different orthopaedic shoe companies and patients. Most shoes had been prescribed to patients with diabetes mellitus, RA, or osteoarthritis, and shoes were not currently in use by the patient.

The test set was constructed in order to resemble daily clinical practice with respect to mean wear-and-tear and to cover the full range of wear-and-tear of the therapeutic shoes. Therefore, a preliminary test set was composed based on a reference study we conducted (15). In this reference study, we found a mean wear-and-tear score of 40.3 ((standard deviation; SD) 18.7; range 2.5–80.3) 6 months after delivery of the shoes in 87 patients with RA. We assumed the mean of 40.3 to be a valid representation of the mean in daily clinical practice, but the SD to be an underestimation, as in the clinic we see the full range from never used to totally worn out. Thus, in order to construct a test set representing wear-and-tear seen in daily clinical practice, we aimed to compose a test set with a mean wear-and-tear of approximately 40.3, a SD greater than 18.7, and including shoes with a minimum score of 0 and maximum score of 100. To assess the representativeness of the preliminary test set, 2 physiatrists first rated all shoes in this test set. Their mean ratings indicated an under-representation of never or seldom worn shoes and extremely worn shoes. As a consequence, the preliminary test set was enriched with never or seldom, and extremely worn shoes. Thus, the final test set included more shoes at the extremes of the range. This final test set was used in the current study.

### *Wear-and-tear scale*

In the wear-and-tear scale (Appendix I), raters are instructed to pay attention to the following details when rating the shoes: wrinkling of the upper, wear-and-tear of laces and Velcro, damage to the leather, especially at the shoe's nose, damage to the outer sole and heel, discoloration and visibility of the footprint, and other signs of wear or repair. Subsequently, raters are asked to rate the degree of wear-and-tear of each therapeutic shoe separately on a 100-mm visual analogue scale (VAS), with 0 and 100 indicating "not at all worn out" and "totally worn out", respectively. The distance (in mm) from 0 to the marked bar of the raters is recorded by a research assistant.

### *Study procedure*

The study design aimed to match clinical practice in the Netherlands. In brief, TF is prescribed, delivered and evaluated at a multidisciplinary foot-care clinic. The patients' visits to the clinic are scheduled over the whole day and the intervals between visits to the multidisciplinary foot-care clinic vary from 3–6 weeks. Both physiatrists and orthopaedic shoe technicians are involved in the foot-care clinic. In both, experience with TF varies considerably.

For the study, raters rated all shoes twice on one day with a 4-h interval (short-term intra-rater test-retest). This procedure was repeated after 4 weeks (long-term intra-rater test-retest), so each rater rated all shoes 4 times. A total of 24 raters participated: 12 physiatrists (including trainees) and 12 orthopaedic shoe technicians; in each group 6 persons were experienced and 6 inexperienced. The raters' skills level was not measured, but their degree of experience was assessed by recording the years of experience at the multidisciplinary foot-care clinic. The median years of experience at a multidisciplinary foot-care clinic was chosen as the cut-off point to distinguish experienced from inexperienced raters.

The numbered boxes with shoes of the test set were presented to the raters in a random order at all 96 (=24 [raters] × 4 [sessions]) measurement sessions. Raters were blinded for their own previous ratings and the results of the other raters. Test conditions (rooms, light and temperature) were standardized. The total number of ratings for this study consisted of 9,600 (=100 [shoes] × 2 [hour] × 2 [week] × 2 [profession] × 2 [experience] × 6 [rater]) ratings.

### *Statistical analysis*

Descriptive statistics (mean (SD)), based on the 24 scores by all raters of each shoe, were used to describe the wear-and-tear of each individual shoe and of all shoes. Independent samples *t*-test were used to compare wear-and-tear scores of all right and all left shoes. Moreover, the wear-and-tear scores within each pair of shoes were compared. In addition, descriptive statistics (median, interquartile range (IQR)) were computed for the years of experience of the involved professional). SPSS, version 20 calculated descriptive statistics (www.spss.com).

Given the multiple potential sources of lack of reliability or error-variance (short-term and long-term test-retest, profession, experience, and rater) generalizability or G-theory was used for this study (16, 17). In G-theory, it is recognized that in any measurement situation there are multiple sources of error variance. Moreover, in the first stage of G-theory, a so-called Generalizability or G-study, the individual error variances are estimated simultaneously. In this G-study all the plausible sour-

**Table I.** Overview of the reliability coefficients of the wear-and-tear scale that were calculated, their interpretation, and the facets of generalizability that were considered to be random and fixed, respectively

| | | Facet of generalizability | | | | |
|---|---|---|---|---|---|---|
| Type of reliability | Interpretation | Hour | Week | Profession | Experience | Rater |
| *Intra-rater or test-retest* | | | | | | |
| Short-term: hour 0 vs hour 4 | Reliability for the same rater with a 4-h interval | Random | Fixed | Fixed | Fixed | Fixed |
| Long-term: week 0 vs week 4 | Reliability for the same rater with a 4-week interval | Fixed | Random | Fixed | Fixed | Fixed |
| Overall | Reliability for the same rater with a 4-h or 4-week interval | Random | Random | Fixed | Fixed | Fixed |
| *Inter-rater* | | | | | | |
| Profession: physician vs shoe technician | Reliability between physicians and shoe technicians | Fixed | Fixed | Random | Fixed | Fixed |
| Experience: ≤4 years vs >4 years | Reliability between inexperienced and experienced raters | Fixed | Fixed | Fixed | Random | Fixed |
| Overall | | Fixed | Fixed | Fixed | Fixed | Random |

**Table II.** Intra-rater and inter-rater reliability coefficients of the wear-and-tear scale and associated standard errors of measurement (SEM) and smallest detectable difference (SDD) or change (SDC)

| Type of reliability | Absolute reliability coefficient | SEM | SDD or SDC |
|---|---|---|---|
| *Intra-rater or test-retest* | | | |
| Short-term: hour 0 vs hour 4 | 0.99 | 2.6 | 7.3 |
| Long-term: week 0 vs week 4 | 0.99 | 2.9 | 8.0 |
| Overall | 0.98 | 3.9 | 10.8 |
| *Inter-rater* | | | |
| Profession: physician vs shoe technician | 0.97 | 4.9 | 13.7 |
| Experience: ≤4 years vs >4 years | 0.98 | 4.5 | 12.4 |
| Overall | 0.93 | 8.1 | 22.5 |

ces of error are incorporated into a single analysis of variance, and the individual variance components are computed. Then, through the second part of G-theory, the Decision or D-study, the impact of the sources of error-variance and various combinations can be examined.

*G-Study.* A 5-facet fully crossed design was used (16). The object of measurement or facet of differentiation was: shoe. The 5 facets of generalization were: hour (hour 0 vs hour 4), and week (week 0 vs week 4), profession (physician vs shoe technician), experience (≤4 years vs >4 years) and rater (rater 1 to rater 6).

*D-Study.* Absolute reliability coefficients (or intraclass correlations coefficients for agreement) (14) of intra-rater, inter-rater and general reliability were calculated by considering different facets of generalization as fixed or random factors (14). Table I gives an overview of the facets of generalization that were considered fixed or random, respectively. G-theory was applied with the software program G_string_IV, version 6.3.7. (18). Reliability coefficients were interpreted as follows: <0.40: poor; 0.40–0.74: fair to good and ≥0.75 excellent (19, 20).

*Sample size.* In generalizability studies there is no feasible strategy to compute sample size (21). Therefore, the sample size was chosen on the basis of convenience and feasibility.

In addition to the reliability coefficients, relative measures of reliability, the standard errors of the measurement (SEMs), absolute measures of reliability, were calculated (17) and from these SEMs the smallest detectable differences (SDDs) (22) or smallest detectable change (SDCs) (20) were derived. The SDD indicates the difference between ratings of different shoes on the measurement scale, and the SDC indicates the change in ratings of the same shoe. This means that these values (SDD and SDC, respectively) numerically represent a real difference or change, i.e. not attributable to measurement error. For this study, the SEMs were calculated as the square-root of the mean square estimate for the error term at issue, determined using G-theory, and the SDDs and SDCs as $1.96 \times \sqrt{2} \times SEM$ (18).

## RESULTS

The mean wear-and-tear score of the shoes was 39.7 (SD 33.3) with 7 shoes with a score <5.0 and 3 shoes with a score >95.0. No significant differences were found between left (mean 39.8; SD 33.7) and right (mean 39.6; SD 32.9) shoes ($p=0.72$). The mean difference of the wear-and-tear scores between the left and right shoe was –0.2 (95% confidence interval (95% CI) –1.6–1.1) and the mean difference between the wear-and-tear scores of the least worn shoes of each pair and most worn shoes was –3.3 (95% CI –4.6–1.9), indicating a minimal difference. Raters had a median (IQR) years of experience of 4.0 (IQR 1–10).

The reliability coefficients, SEMs, SDDs and SDCs are summarized in Table II. The short-term, long-term and overall intra-rater or test-retest reliability coefficients were 0.98–0.99, which can be interpreted as excellent. The related SEMs ranged from 2.6 to 3.9, and SDDs/SDCs from 7.3 to 10.8. The inter-rater reliability coefficients were 0.97 to 0.98, which can be interpreted as excellent. The related SEMs ranged from 4.5 to 4.9, and SDDs/SDCs from 12.4 to 13.7. The overall reliability or ability to generalize from a single rating was 0.93, which can be interpreted as excellent. The related SEM and SDC/SDD were 8.1 and 22.5, respectively. For researchers who want to calculate their own coefficients, SEMs, and SDDs/SDCs, the underlying variance components are shown in Appendix SI[1]. An overview of different wear-and-tear scores is shown in Appendix SII[1]

## DISCUSSION

The results of this generalizability study show that the wear-and-tear scale is a reliable instrument to measure the wear-and-tear of TF, with an excellent intra-rater and inter-rater reliability and an excellent overall reliability.

For the G-Study we carefully composed our test set. This study included TF with a mean wear-and-tear score comparable to the mean wear-and-tear found in a recent reference study at 6 months after delivery of TF in patients with RA (15). However, the SD of the mean wear-and-tear score of the test set used in this study was intentionally higher than the SD in the reference study. A higher SD will result in a higher variance with respect to the wear-and-tear in the test set. Higher variances result in better (higher) reliability coefficients. As the variance seen in daily clinical practice is expected to be lower, the reliability of the scale in practice will also be lower. Therefore, for future research we recommend continued validation of the wear-and-tear scale. For instance, through collecting new ratings of TF directly, 6 months and 1 year after delivery, in order to determine whether our test set consisted of shoes with a wear-and-tear comparable to daily clinical practice.

The test set was assembled from shoes of patients with diseases in which the feet are generally symmetrically affected. This is reflected by the small difference between the wear-and-tear scores of the least and most

[1]http://www.medicaljournals.se/jrm/content/?doi=10.2340/16501977-2339

worn shoes of each pair. In our view, generalization to patients with asymmetrically affected feet, e.g. patients after stroke, is possible because we investigated the reliability separately for each shoe. Further research is necessary, to investigate the differences in the reliability of the wear-and-tear scale of the least and most worn shoes. In the meantime we recommend rating the most worn shoe in asymmetrical cases.

For the G-study we carefully chose our facets of differentiation. The chosen time intervals were realistic and relevant to daily clinical practice. The included professions were relevant for the multidisciplinary foot-care clinic that participated in our study. Moreover, the full range of experience of raters was covered for the setting of our multidisciplinary foot-care clinic. Finally, the results of this study are based on a large number of raters. For future research we recommend including other prescribers of TF, e.g. orthopaedic surgeons, as raters.

In the D-study the different reliability coefficients were calculated. All reliability coefficients were above 0.93. This is relatively high in comparison with other studies (20). In the calculation of reliability coefficients the variance in the object of measurement is divided by the variance in the object of measurement plus the error variance (17, 20). Consequently, a relatively high variance in the object of measurement, as was intentionally the case in this study, will result in high reliability coefficients. Moreover, the error variance in this study was, by applying generalizability theory, attributed to 5 different sources of variation. In the calculation of specific reliability coefficients, only the error variance that is attributable to the source of variation at issue is included. In our view, this reflects one of the great advantages of generalizability theory.

The short-term and long-term intra-rater reliability are more or less comparable, with coefficients of 0.99, indicating that ratings by the same person are as reliable when separated by a short period of time compared with a long period of time. Possible explanations for these comparable coefficients are the large number of shoes in the test set, with a small chance of recall of previous ratings, even in case of repeated ratings on the same day. Moreover, when stored in a box under stable conditions, the wear-and-tear of the therapeutic shoe did not change over time. In other reliability studies there is often an issue that reliability has to be studied in "stable patients" and, related to this, if the reliability is poor, to differentiate between the lack of reliability of the instrument vs the lack of stability of the phenomenon at issue in the patients under investigation (20). Clearly, this issue does not apply in the current study because the wear-and-tear was studied in "stable shoes", which may result in reliability coefficients that are relatively high in comparison with other reliability studies. Given the mean wear-and-tear score (40.3) in our reference study at 6 months after delivery of TF, the SDCs (7.3–10.8) are small enough to detect change over time if the wear-and-tear scale is applied by one rater only.

The inter-rater reliabilities and overall reliability coefficients were 0.97, 0.98 and 0.93, respectively. The SDCs are relatively high (12.4–22.5). However, they are still sufficiently small enough for the application of the wear-and-tear scale by different raters in clinical practice, if we assume that the wear-and-tear will increase linearly, in the time interval from 6 months up to one year after delivery of the shoes, even above 40.3. However, this assumption should be confirmed in future research covering the full-time interval up to one year after delivery. For the application of the wear-and-tear scale in daily clinical practice, i.e. in the setting of a multidisciplinary foot-care clinic, and in order to detect smaller changes, we recommend using the scores of one rater over time or the mean scores of 2 different raters. The mean scores of 2 ratings, by an inexperienced or experienced physiatrist or orthopaedic shoe technician, result in a reliability coefficient of 0.96 and a SDC of 15.9.

There is a limited, although greater, consistency between ratings made by the same rater (0.99) compared with ratings made by different raters (0.97–0.98). This might be explained by the fact that each rater makes their own weighting of the different indicators of wear-and-tear observed on the shoes separately and/or together.

The present study found that the wear-and-tear scale is a reliable indicator of wear-and-tear of TF. For its application in daily clinical practice, we chose one global judgment with 1 VAS score instead of 4 separate VAS scores for the different parts of the shoe, which gives detailed information, but is rather time-consuming. The wear-and-tear scale may be used by TF prescribers and professionals involved in TF prescription to endorse the replacement of TF with a new pair, although threshold values for replacement are not yet available. Future studies should focus on threshold values for replacement of TF.

With respect to application of the wear-and-tear scale in daily clinical practice, the question arises as to whether the wear-and-tear scale might also be used by TF prescribers and professionals involved in the prescription of TF, as an indicator of TF use. Studies have shown that TF use is suboptimal, with varying non-use rates of 17–25% (15, 23), thus reducing its effectiveness (24). Several instruments are available to assess TF use. Subjective instruments comprise patient-reported measurement instruments (25) and objective instruments are mostly performance-based (26). They have disadvantages, such as the potential for response bias, missing data, not being very feasible in

daily clinical care, and being costly. The wear-and-tear scale tackles these downsides: it is a simple, rapid, clinically easily applicable and inexpensive measurement instrument. However, a drawback of the wear-and-tear scale is that the extent of wear-and-tear of the shoes may also be influenced by the patient's weight, asymmetrical gait patterns, lower limb (mal)alignment, the applied shoe materials, walking surfaces and climate or weather conditions. Thus, the wear-and-tear scale is not a pure indicator of TF use. For future research we recommend studying the validity of the wear-and-tear scale as an indicator of TF use by comparing the scale with currently existing indicators of use. In these studies associations between wear-and-tear and use, will have to be corrected for potential confounders, such as patient's weight, asymmetrical gait patterns, shoe materials applied, and climate or weather conditions.

### *Study strengths and limitations*

The strengths of this study are its design, with a carefully assembled test set with a large number of therapeutic shoes and a large number of raters, combined with application of the G-theory. According to the COSMIN (COnsensus-based Standards for the selection of health status Measurement INstruments) criteria the methodological quality of this study is excellent because no ratings were missing, the sample size was adequate (100), more than 2 measurements were available, the administrations were independent, the time interval was appropriate (4 weeks), the object of measurement, the wear-and-tear of the therapeutic shoe, did not change over time, test conditions were similar for all measurements, there were, to our knowledge, no important flaws in the design of methods of the study, and intraclass correlation coefficients for agreement were calculated (27).

A limitation of the current study is that it was conducted in the Netherlands and therefore may not reflect clinical practice in other countries. Furthermore, there was no detailed information about the wear-and-tear of all shoes in the interval between taking off the shoes for the last time and commencement of the study, and shoes of patients with diseases with asymmetrically affected feet were not included. Future studies into the reliability of the wear-and-tear scale in other clinical settings and in patients with asymmetrically affected feet are needed.

### *Conclusion*

This generalizability study, with both a large number of therapeutic shoes and raters, showed that the wear-and-tear scale is a feasible, simple and rapid measurement instrument that can be used to reliably measure the wear-and-tear of TF.

## ACKNOWLEDGEMENT

The authors thank José de Vries for her support with the logistics of this study.

*The authors have no conflicts of interest to declare.*

## REFERENCES

1. Riskowski JL, Dufour AB, Hagedorn TJ, Hillstrom HJ, Casey VA, Hannan MT. Associations of foot posture and function to lower extremity pain: results from a population-based foot study. Arthritis Care Res 2013; 65: 1804–1812.
2. Boulton AJ, Kirsner RS, Vileikyte L. Clinical practice. Neuropathic diabetic foot ulcers. N Engl J Med 2004; 351: 48–55.
3. Otter SJ, Lucas K, Springett K, Moore A, Davies K, Cheek L, et al. Foot pain in rheumatoid arthritis prevalence, risk factors and management: an epidemiological study. Clin Rheumatol 2010; 29: 255–271.
4. Hill CL, Gill TK, Menz HB, Taylor AW. Prevalence and correlates of foot pain in a population-based study: the North West Adelaide health study. J Foot Ankle Res 2008; 1: 2.
5. Chalmers AC, Busby C, Goyert J, Porter B, Schulzer M. Metatarsalgia and rheumatoid arthritis – a randomized, single blind, sequential trial comparing 2 types of foot orthoses and supportive shoes. J Rheumatol 2000; 27: 1643–1647.
6. Michelson J, Easley M, Wigley FM, Hellmann D. Foot and ankle problems in rheumatoid arthritis. Foot Ankle Int 1994; 15: 608–613.
7. Allet L, Armand S, Golay A, Monnin D, de Bie RA, de Bruin ED. Gait characteristics of diabetic patients: a systematic review. Diabetes Metab Res Rev 2008; 24: 173–191.
8. Grondal L, Tengstrand B, Nordmark B, Wretenberg P, Stark A. The foot: still the most important reason for walking incapacity in rheumatoid arthritis: distribution of symptomatic joints in 1,000 RA patients. Acta Orthopaedica 2008; 79: 257–261.
9. Hogg FR, Peach G, Price P, Thompson MM, Hinchliffe RJ. Measures of health-related quality of life in diabetes-related foot disease: a systematic review. Diabetologia 2012; 55: 552–565.
10. Turner DE, Helliwell PS, Siegel KL, Woodburn J. Biomechanics of the foot in rheumatoid arthritis: identifying abnormal function and the factors associated with localised disease 'impact'. Clin Biomech (Bristol, Avon) 2008; 23: 93–100.
11. World Health Organization. ICF: International Classification of Functioning, Disability and Health. ICF. Geneva: WHO; 2001. Available from: http://www.who.int/classification/icf 2001.
12. Farrow SJ, Kingsley GH, Scott DL. Interventions for foot disease in rheumatoid arthritis: a systematic review. Arthritis Rheum 2005; 53: 593–602.
13. Albrecht K, Richter A, Callhoff J, Huscher D, Schett G, Strangfeld A, et al. Body mass index distribution in rheumatoid arthritis: a collaborative analysis from three large German rheumatoid arthritis databases. Arthritis Res Ther 2016; 18: 149.
14. Vet de HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol 2006; 59: 1033–1039.
15. Dahmen R, Buijsmann S, Siemonsma PC, Boers M, Lankhorst GJ, Roorda LD. Use and effects of custom-made therapeutic footwear on lower-extremity-related pain and activity limitations in patients with rheumatoid arthritis: a prospective observational study of a cohort. J Rehabil Med 2014; 46: 561–567.
16. Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. Med Teach 2012; 34: 960–992.

17. Vidal C BT, Morel J, Combe B, Daïen C. Association of body mass index categories with disease activity and radiographic joint damage in rheumatoid arthritis: a systematic review and metaanalysis. J Rheumatol 2015; 42: 2261–2269.
18. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham CO, 3rd, et al. 2010 Rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. Arthritis Rheum 2010; 62: 2569–2581.
19. Platto MJ, O'Connell PG, Hicks JE, Gerber LH. The relationship of pain and deformity of the rheumatoid foot to gait and an index of functional ambulation. J Rheumatol 1991; 18: 38–43.
20. Vet de HC, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine: a practical guide. Cambridge: Cambridge University Press; 2011.
21. Brennan RL. Generalizability theory. Educational Measurement: Issues and Practice 1992; 11: 27–34.
22. Beckerman H, Roebroeck ME, Lankhorst GJ, Becher JG, Bezemer PD, Verbeek AL. Smallest real difference, a link between reproducibility and responsiveness. Qual Life Res 2001; 10: 571–578.
23. Jannink MJ, Ijzerman MJ, Groothuis-Oudshoorn K, Stewart RE, Groothoff JW, Lankhorst GJ. Use of orthopedic shoes in patients with degenerative disorders of the foot. Arch Phys Med Rehabil 2005; 86: 687–692.
24. Bus SA, Waaijman R, Arts M, de Haart M, Busch-Westbroek T, van Baal J, et al. Effect of custom-made footwear on foot ulcer recurrence in diabetes: a multicenter randomized controlled trial. Diabetes Care 2013; 36: 4109–4116.
25. Netten van JJ, Hijmans JM, Jannink MJ, Geertzen JH, Postema K. Development and reproducibility of a short questionnaire to measure use and usability of custom-made orthopaedic shoes. J Rehabil Med 2009; 41: 913–918.
26. Armstrong DG, Lavery LA, Kimbriel HR, Nixon BP, Boulton AJ. Activity patterns of patients with diabetic foot ulceration: patients with active ulceration may not adhere to a standard pressure off-loading regimen. Diabetes Care 2003; 26: 2595–2597.
27. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res 2012; 21 651–657.

**Appendix I**. Wear-and-tear scale.

**Instructions**

Would you please be so kind to give your opinion on the extent to which the shoe is worn? Before this, please read through the attention points carefully. The attention points consist of wearing characteristics of the different parts of the shoe. Now, formulate your opinion on the extent of the wearing out of the shoe by means of marking a bar on the horizontal line.

**Attention points**

- wrinkling of the upper
- wear-and-tear of laces and Velcro
- damage of the leather, especially at the shoe's nose
- damage of the outer sole and heel
- discoloration and visibility of the foot print
- other signs of wearing or repair

**Example**

A totally worn out shoe as is indicated on the line as follows:


**Your opinion**

Mark with a | on the line the extent to which the shoe is worn:

Right shoe:


Left shoe: