Validating the 28-Tender Joint Count Using Item Response Theory

LISETH SIEMONS, PETER M. ten KLOOSTER, ERIK TAAL, INA H. KUPER, PIET L.C.M. van RIEL, MART A.F.J. van de LAAR, and CEES A.W. GLAS

ABSTRACT. ***Objective.*** To examine the construct validity of the 28-tender joint count (TJC-28) using item response theory (IRT)-based methods.

Methods. A total of 457 patients with early stage rheumatoid arthritis (RA) were included. Internal construct validity of the TJC-28 was evaluated by determining whether the TJC-28 fit a 2-measure logistic IRT model. As well, we tested whether the discrimination and difficulty parameters of the joints properly reflected the known left-right symmetry of joint involvement. External validity was evaluated by correlations with other established measures of disease activity, including pain, disability, general health, erythrocyte sedimentation rate (ESR), and the 28-swollen joint count.

Results. The TJC-28 showed a good fit with the 2-parameter logistic model, with no relevant differential item functioning across sex, age, and time and with excellent reliability. The 28 joints covered a reasonable range of disease activity, even though they were mainly targeted at patients with moderate or high disease activity levels. The joint parameters reflected the left-right symmetry of joint involvement for all pairs of joints except one. All disease activity measures, except ESR, were significantly correlated with the TJC-28. Most correlations were of the expected magnitude.

Conclusion. The TJC-28 showed good internal and acceptable external construct validity for patients with early-stage RA. The IRT analyses did point to some potential limitations of the instrument, a major problem being its limited measurement range. Future research should examine whether instrument modifications might lead to a more robust assessment of disease activity in patients with RA. (J Rheumatol First Release Oct 1 2011; doi:10.3899/jrheum.110436)

Key Indexing Terms:
2-PARAMETER LOGISTIC MODEL 28-TENDER JOINT COUNT COHORT STUDY
CONSTRUCT VALIDITY ITEM RESPONSE THEORY RHEUMATOID ARTHRITIS

Rheumatoid arthritis (RA) is a systemic autoimmune disease, decreasing life expectancy by 3 to 10 years compared to the general population[1,2]. People with RA experience chronic inflammation of joints and periarticular tissues[3], characterized by symmetric pain and swelling in the joints[4,5,6,7]. The disease generally follows an unpredictable course, often with alternating periods of mild and severe disease activity[3].

RA treatments are aimed at reaching a state of remission as soon as possible[8]. Because joint tenderness is an important characteristic of RA, joint counts that measure the extent of joint tenderness are used for the assessment of RA severity[9]. A joint count is a specific quantitative clinical measure to assess the status of a patient with RA[10]. Therefore, it forms a major component of indices of disease activity[11] and remission[12]. Although various joint counts have been developed, ranging from the evaluation of 28 to 80 joints, the 28-joint count is currently the most widely used measurement instrument.

Earlier studies showed the 28-tender joint count (TJC-28) to be a reliable and valid joint index[9,13,14,15]. However, these studies have used only classical test theory (CTT) methods. To date, the construct validity of the TJC-28 has never been analyzed using item response theory (IRT)-based methods. IRT is a sophisticated psychometric approach that has been adopted to supplement the more traditional approaches[16] to enable a more thorough evaluation of an instrument's psychometric characteristics. IRT has already

From the Arthritis Center Twente, Department of Psychology, Health and Technology, and the Department of Rheumatology, Medisch Spectrum Twente, and the Department of Research Methodology, Measurement and Data Analysis, University of Twente, Enschede; and Department of Rheumatology, University Medical Centre St. Radboud, Nijmegen, The Netherlands.

L. Siemons, MSc; P.M. ten Klooster, PhD; E. Taal, PhD, Arthritis Center Twente, Department of Psychology, Health and Technology; University of Twente; I.H. Kuper, MD, PhD, Arthritis Center Twente, Department of Rheumatology, Medisch Spectrum Twente; P.L.C.M. van Riel, MD, PhD, Department of Rheumatology, University Medical Centre St. Radboud; M.A.F.J. van de Laar, MD, PhD, Arthritis Center Twente, Department of Psychology, Health and Technology, University of Twente, Department of Rheumatology, Medisch Spectrum Twente; C.A.W. Glas, PhD, Arthritis Center Twente, Department of Research Methodology, Measurement and Data Analysis, University of Twente.

Address correspondence to L. Siemons, Faculty of Behavioural Sciences, Department of Psychology, Health and Technology, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands.
E-mail: l.siemons@utwente.nl

Accepted for publication July 29, 2011.

been frequently and successfully applied in evaluating and improving health outcome questionnaires[17], but it has rarely been applied to clinical measures, such as tender joint counts. Therefore, the aim of our study was to examine both the internal and external construct validity of the TJC-28 using IRT-based methods.

MATERIALS AND METHODS

Patients with early-stage RA participating in the Dutch Rheumatoid Arthritis Monitoring remission induction cohort[18] were included in this study. This observational, multicenter cohort was established in 2006 to evaluate a treatment strategy aimed at reaching a state of remission. The patients were asked for inclusion in the cohort by their rheumatologists. Patients were qualified for inclusion at the moment of clinical diagnosis of RA. Symptom duration was a maximum of 1 year, and patients had to be at least 18 years old. Any who had previously used disease-modifying antirheumatic drugs or prednisolone were excluded from the cohort.

The result was a total baseline sample of 457 patients. Measurements were performed during each hospital visit. The data from the first timepoint (i.e., at inclusion) were used for all analyses. In addition, one of the fit analyses (i.e., evaluating differential item functioning across time) was based on data from the first 3 timepoints (t1 = at inclusion, t2 = 8 weeks after inclusion, t3 = 12 weeks after inclusion). Because the duration since inclusion varied among patients, followup measurements involved a decreasing number of patients. At the third timepoint, the remaining sample consisted of a total of 391 patients.

Measures. The TJC-28 and the 28-swollen joint count (SJC-28) were administered separately at each visit by a trained nurse practitioner or rheumatologist. The 28 joints were scored on a dichotomous scale, with 0 indicating "no pain" or "no swelling" in the joint, and 1 indicating "pain" or "swelling" in the joint[9,19]. Both 28-joint counts include the shoulders, elbows, wrists, and knees, the 10 metacarpophalangeal (MCP) joints, and the 10 proximal interphalangeal (PIP) joints[20].

Besides the TJC-28 and the SJC-28, patients were asked to complete the Health Assessment Questionnaire-Disability Index (HAQ-DI)[21], which measures physical function, and visual analog scales for pain (VAS pain) and general health (VAS GH). The alternative disability index (HAQ-ADI)[22], which does not correct for the use of aids and devices, was derived from the HAQ-DI and was scored on a scale from 0 to 3 (higher scores indicating more physical disability). Pain and general health were measured using a 100-mm VAS scale, 0 indicating "no pain" or "very good", and 100 indicating "unbearable pain" or "very bad."

Laboratory samples were collected before each hospital visit, including the erythrocyte sedimentation rate (ESR), which is a nonspecific measure of inflammation[19].

Statistical analyses. When using an IRT framework, the relationship between item scores and the underlying construct of interest (i.e., the latent trait variable θ, representing the degree of joint tenderness in our study) can be modeled. When applying CTT approaches, sum scores of different combinations of joints can be obtained. However, this does not imply that the sum score reflects a meaningful underlying construct. IRT has several beneficial properties compared to the traditional CTT approach, enabling a more thorough evaluation of an instrument's psychometric characteristics. If the TJC-28 fits an IRT model, this supports the construct validity of the instrument, because this shows that the observed responses can be explained by the underlying structure of the instrument[23]. Further, if attenuation is present (i.e., underestimated correlations between measurements due to unreliability caused by measurement error), IRT can deal with this problem more precisely than the CTT approach since it considers latent correlations instead of sum-score based observed correlations. In addition, IRT can successfully handle incomplete item administration designs and missing data, and where CTT often assumes a normal distribution of the true scores, IRT can deal with various distributions of latent variables[23].

IRT models the probability of a joint being scored as tender on the basis of characteristics of the patient (the degree of joint tenderness: θ) and the item (such as the difficulty and discrimination level). Each single joint is regarded as an item and has a corresponding IRT model curve. Two widely applied IRT models are the Rasch model (also known as the 1-parameter logistic model) and the 2-parameter logistic (2-PL) model, both shown in Figure 1. The y-axis shows the probability of a joint to be scored as tender, while the x-axis shows the latent trait that corresponds to the degree of joint tenderness a patient experiences (θ, scaled around zero). Figure 1A shows the Rasch model, including 3 joints with different difficulty parameters[24]. The value of the difficulty parameter of a specific joint equals the point on the x-axis at which the patient has a probability of 0.5 of having a painful joint[24,25]. So for joint 1, its value will be equal to –1. Figure 1B shows the 2-PL model. In it, the curves intersect because of addition of the discrimination parameter. This parameter is proportional to the slope of the curve; the higher its value, the steeper the slope, and the better the joint discriminates between patients with various degrees of joint tenderness[25].

In our study, a 2-PL model was used to analyze the construct validity of the TJC-28. This was motivated by both practical and empirical reasons. First, we wanted to examine whether the symmetry that characterizes RA is reflected in both the difficulty and the discrimination parameters of the IRT model. Second, a log-likelihood ratio test showed that the 2-PL model had a significantly better fit to the TJC-28 than the Rasch model (log-likelihood ratio test = 163.81, df = 27, $p < 0.01$).

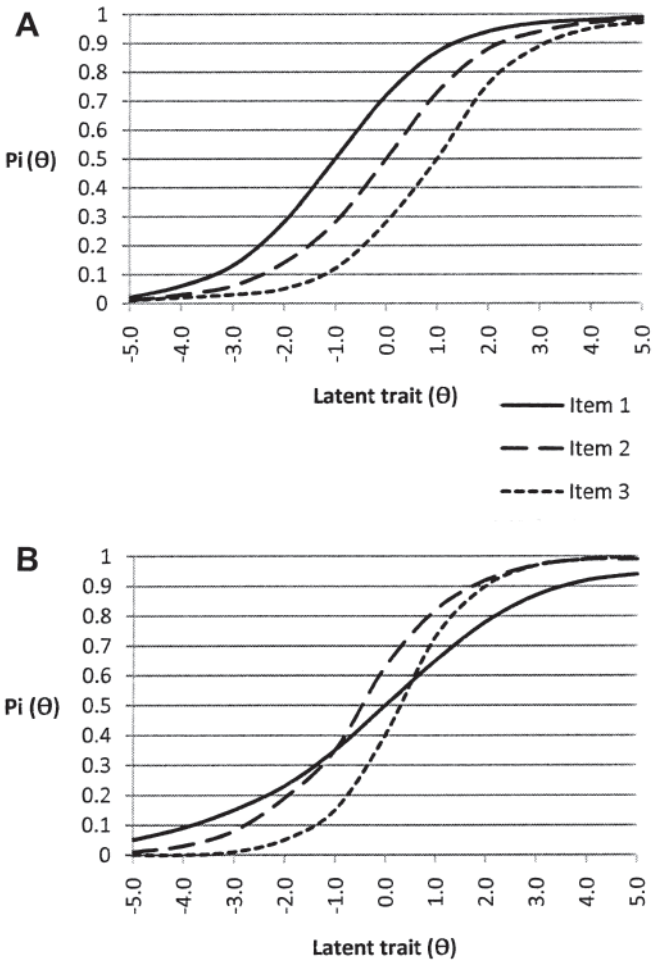

Figure 1. Graphic representation of the Rasch model (A) and the 2-parameter logistic model (B), where Pi(θ) is the probability of a joint being scored as tender.

Internal construct validity. This was assessed by evaluating whether the TJC-28 could be fitted to the 2-PL model, whether the joint parameters truly reflected the known left-right symmetry of joint involvement in patients with RA[4,5], and whether the TJC-28 had an acceptable reliability.

Fit analyses. IRT models rely on several assumptions. One of these concerns the shape of the response curves. Using a Lagrange multiplier test, the LM-Q1-test[26], it was determined whether the shape of the curves belonging to the TJC-28 fit the shape of the curves assumed by the 2-PL model. This means the joint curves have various difficulty parameters, various discrimination parameters, and a lower zero asymptote. Two outcome values considered important for determining the fit of the curves with the LM-Q1-test are the p value of the test and the effect size[27]. A p value > 0.05 indicates a good item-model fit, but this statistic is sensitive to large sample sizes[27]. For large sample sizes, the absolute effect size should also be evaluated. The effect size is given by the difference between the observed and the expected average score on an item in a specific group and can, therefore, range between 0 and 1. An effect size of < 0.10 has been previously used as an acceptable measure for item model fit[28]. Well-fitting response curves can also be seen as strong evidence for unidimensionality of the TJC-28[29].

Additionally, it was examined whether differential item functioning (DIF) across sex, age, and time were present. A joint shows DIF across sex or age if individuals from different groups (e.g., men vs women) but with the same latent trait value do not have the same probability of reporting a joint as being tender[24]. DIF across time is present when the joint difficulty parameters are unstable over time[27]. The stability of the parameters was examined over the first 3 timepoints.

Left-right symmetry of joint parameters. The symmetry of the difficulty and discrimination parameters was simultaneously tested for each pair of joints using a Wald test[30]. This test determines whether the parameter values of the left-side joint and the parameter values of the right-side joint are equal. Nonsignificant results (p > 0.05) indicate that the joint parameters properly reflect the known left-right symmetry of joint involvement.

Reliability and measurement precision. In IRT, the reliability of the TJC-28 is estimated as the ratio of the expectation of the posterior variance of the latent variable θ given the instrument score, and the total variance of θ[29]. A θ value > 0.70 is considered acceptable for group use, while a value of 0.85 or higher is required for individual use[16]. The IRT reliability coefficient is equivalent to Cronbach's alpha.

When applying IRT, the range of θ for which a joint or the total TJC-28 is most reliable for measuring patients' levels of joint tenderness can be depicted in an information curve. An information curve shows the range over θ where the individual joint or the total TJC-28 can best discriminate among individual patients[31]. Ideally, the instrument includes joints with high discrimination parameters that cover a broad spectrum of joint difficulties. In this way, the spectrum of joint tenderness can be measured as precisely as possible. The higher the information level of a joint, the more the joint contributes to the measurement precision of joint tenderness. Information curves of individual joints were plotted for evaluation of the performance of each single joint. The test information curve of the TJC-28 and its associated reliability levels [r = 1 – (1/test information at θ)] were plotted to evaluate the performance of the total TJC-28.

External construct validity. Previous studies used sum scores of the TJC to determine its correlation with other established measures of disease activity, while IRT uses latent trait values (θ). The external construct validity of the TJC-28 was evaluated by examining whether the baseline θ values and traditional sum scores of the TJC-28 showed an expected pattern of correlations with 5 other established measures of disease activity[32]: VAS pain, HAQ-ADI, VAS GH, ESR, and the SJC-28.

Correlations < 0.3 were defined as weak (low), between 0.3 and 0.6 as moderate, and > 0.6 as strong (high)[33]. All correlations were expected to be both positive and significant. Although highly variable correlations between the TJC and these variables were found in previous studies, moderate correlations were expected since they are all measures of disease activity[9,14,34,35,36,37,38].

RESULTS

Demographics at inclusion. Baseline data were available from 457 patients (288 women and 169 men). The mean (SD) age at inclusion was 55.4 (15.2) years for the women and 59.8 (12.4) years for the men. Baseline measures of disease activity are summarized in Table 1. The TJC-28 had a mean score of 5.7. For interpretation, a TJC-28 score of 0 corresponded to an estimated θ score in the range of –1.65 to –0.69, and a TJC-28 score of 28 corresponded to estimated θ scores in the range of 2.82 to 3.25.

Internal construct validity. Table 2 presents the results of the fit analyses. Although some joints showed a statistically significant misfit (p < 0.05), all effect sizes were well below 0.10. These results indicate that there was a good fit between the curves of the TJC-28 and the 2-PL model. In addition, there was no relevant DIF across sex, age (median split: ≤ 59 vs ≥ 60 years), and time.

Left-right symmetry of joint parameters. Table 3 presents the parameter estimates generated by the 2-PL model. The Wald test showed a nonsignificant result for all pairs of joints except 1. This demonstrates that both the difficulty and the discrimination parameters properly reflected the left-right symmetry of joint involvement, which is characteristic of RA.

Reliability and measurement precision. The reliability of the TJC-28 was acceptable for group use as well as for individual use (r = 0.874).

Table 3 presents the discrimination parameter values, ranging from 0.670 to 1.049 for larger joints (shoulders, elbows, wrists, knees), and from 1.369 to 2.269 for smaller

Table 1. Mean scores (SD) of established measures of disease activity at baseline in 457 patients with early-stage rheumatoid arthritis.

Measures	Scoring Scale	Mean (SD)
TJC-28	0–28	5.7 (5.7)
VAS pain	0–100	49.4 (25.4)
HAQ-ADI	0–3	1.0 (0.7)
VAS GH	0–100	49.9 (25.2)
SJC-28	0–28	7.9 (5.7)
ESR	0–140	29.6 (22.0)
DAS28	0–10	4.7 (1.4)

TJC-28: tender joint count for 28 joints; VAS: visual analog scale; HAQ-ADI: Health Assessment Questionnaire-Alternative Disability Index; GH: patient's general health assessment; SJC-28: swollen joint count for 28 joints; ESR: erythrocyte sedimentation rate; DAS28: Disease Activity Score for 28 joints.

Table 2. Results of the fit analyses.

Fit Analysis	No. Joints with p ≤ 0.05	Effect Size
Fit of the curves	4	≤ 0.03
Sex differences	4	≤ 0.06
Age differences	1	≤ 0.03
Constancy of location parameters over time	7	≤ 0.06

Table 3. Average joint scores, item response theory joint parameter values, and Wald test results.

Joint	Average Joint Score*		Discrimination Parameter		Difficulty Parameter		Results	
	Left	Right	Left	Right	Left	Right	Wald Test**	p†
Shoulder	0.24	0.22	0.721	0.686	1.252	1.350	0.139	0.93
Elbow	0.12	0.13	0.713	0.670	2.212	2.101	0.187	0.91
Wrist	0.35	0.38	0.983	1.094	0.747	0.613	0.295	0.86
MCP1	0.17	0.21	1.453	1.443	2.130	1.796	2.499	0.29
MCP2	0.21	0.25	1.627	1.735	1.867	1.617	1.756	0.42
MCP3	0.18	0.24	1.850	1.943	2.336	1.822	7.056	0.03
MCP4	0.12	0.14	2.269	1.953	3.406	2.804	4.415	0.11
MCP5	0.12	0.11	2.135	2.266	3.255	3.659	2.580	0.28
PIP1	0.13	0.14	1.552	1.369	2.613	2.334	0.974	0.61
PIP2	0.26	0.26	1.554	1.669	1.415	1.469	0.071	0.96
PIP3	0.26	0.33	1.898	1.784	1.650	0.986	4.617	0.10
PIP4	0.21	0.26	1.528	1.577	1.828	1.541	1.328	0.51
PIP5	0.18	0.19	1.579	1.708	2.131	2.196	0.072	0.96
Knee	0.20	0.20	0.778	0.710	1.572	1.516	0.017	0.99

* Average score on a scale from 0 to 1, ** with 2 degrees of freedom; † p value for a simultaneous test for differences in difficulty and/or discrimination. MCP: metacarpophalangeal; PIP: proximal interphalangeal.

joints (the MCP and PIP). Joint difficulties covered only the positive half of the spectrum, ranging from 0.613 to 3.659, reflecting low response probabilities. This limited range of joint difficulties was also reflected in the information curves (Figure 2). The test information curve showed that the scale measured the patient's level of θ with a reliability level acceptable for group use ($r > 0.70$) over the range from $\theta = 0.60$ to $\theta = +3.05$[31]. Outside this range, the test information curve and the scale's reliability rapidly decreased, meaning that the corresponding levels of θ were estimated with reduced precision. Over the range from $\theta = 0.0$ to $\theta = +2.5$, the scale's reliability was also acceptable for individual use ($r > 0.85$). The reliability was at its highest point ($r > 0.93$) at $\theta = +1.3$. The item information curves showed that smaller joints (MCP and PIP) provided more information to the test than larger joints (shoulders, elbows, wrists, knees).

External construct validity. Spearman's correlations with the other established measures of disease activity for both the θ estimations and the sum scores of the TJC-28 are shown in Table 4. The correlations based on the θ estimations of the TJC-28 were very similar to the correlations based on the sum scores. As expected, all correlations were positive. However, for both the θ estimates and the sum scores of the TJC-28, only 4 out of 5 correlations were significant. The HAQ-ADI, joint swelling, and the patient's general health assessment did show the expected moderate correlations. Pain correlated less strongly with joint tenderness than expected, but the correlation was only just below the cutoff point of 0.30. However, a very low correlation was found with ESR.

DISCUSSION

This is the first study to examine the validity of the TJC-28 by applying IRT-based methods. As a result, the instrument's psychometric characteristics can be evaluated more thoroughly than with CTT alone. The results showed that the TJC-28 is a valid and reliable measure for patients with early-stage RA. An acceptable fit of the TJC-28 to the 2-PL model was demonstrated, with no relevant DIF across sex, age, and time, and with excellent reliability. The joints included in the TJC-28 covered a reasonable range of disease activity, although measurement precision was limited for lower levels of disease activity. Additionally, the joint parameters properly reflected the left-right symmetry of joint involvement. Evaluation of the external validity showed that all correlations, except with ESR, were similar to the correlations found in previous studies.

Statistical transformations of the ESR values, such as square root and natural logarithm transformations as performed in the Disease Activity Score for 28 joints[15], did not improve the correlation with joint tenderness. A limited distribution of ESR values within the patient sample might explain the nonsignificance of this correlation. However, given the high SD (22.04) of the ESR values, this does not seem plausible. Moreover, secondary analyses did show significant and higher correlations between ESR and all other measures of disease activity (r between 0.17 for the VAS GH and 0.30 for the HAQ-ADI) and C-reactive protein ($r = 0.64$), another measure of inflammation. Evaluation of the correlations with the individual joints showed that ESR was significantly correlated with the larger joints (r between 0.10 and 0.14), but not with the smaller joints that constitute the largest part of the TJC-28. This higher correlation with larger joints is in accord with earlier findings[37] and suggests that the ESR mainly reflects the volume of inflammation in the larger joints, while the TJC-28 is also in large part explained by the smaller joints. Future studies should evaluate the correlation between the TJC-28 and ESR in an RA population

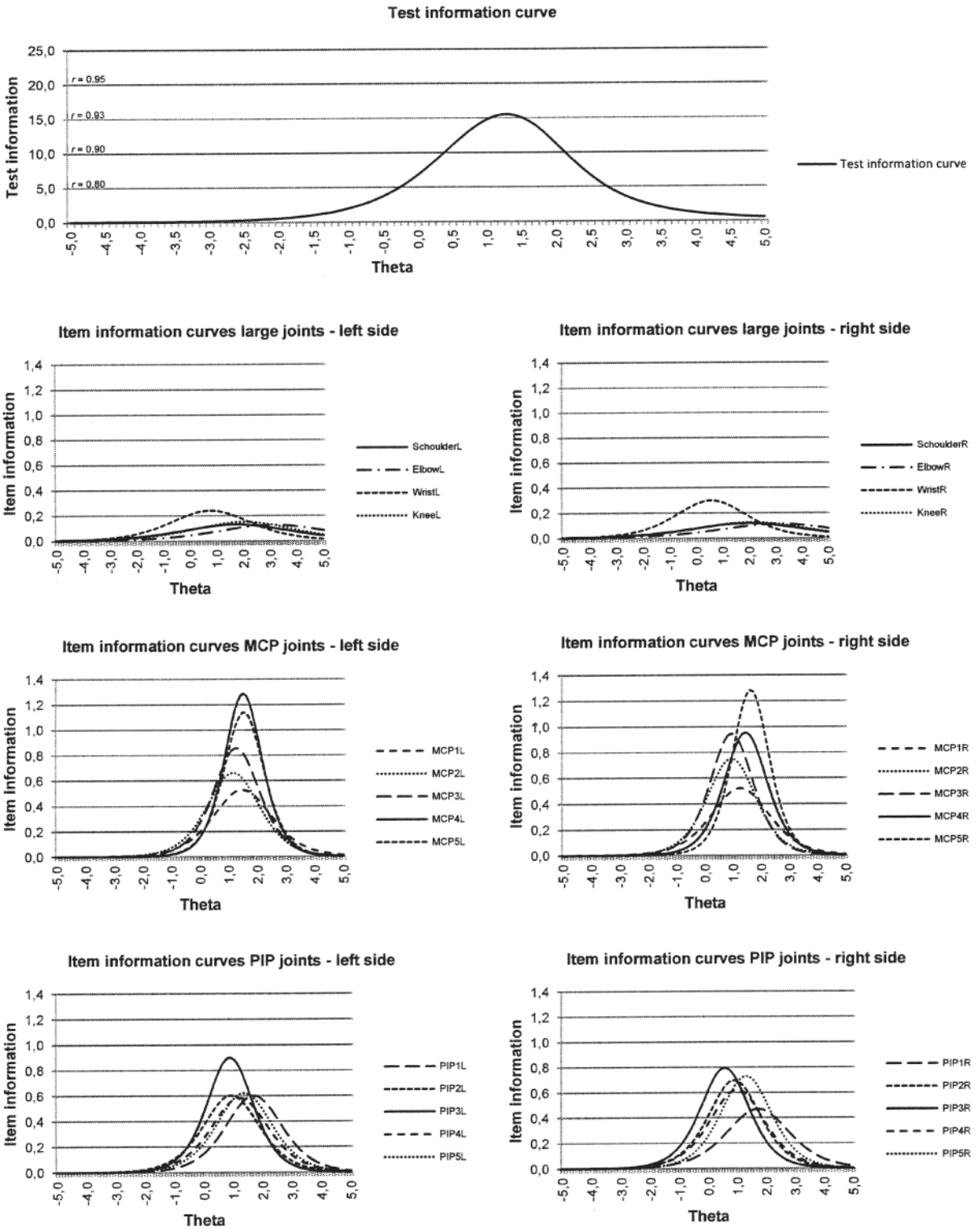

Figure 2. Top graph shows the test information curve of the 28-tender joint count with its associated reliability level. The graphs below represent the item information curves for the joints on the left side (left column) and right side (right column) of the body.

in which the patients have more inflamed smaller joints than in the current sample, to determine whether joint size affects this correlation.

Examination of the correlations showed that those based on the θ estimations of the TJC-28 were very similar to the correlations based on the sum scores of the TJC-28. This

Table 4. Spearman's correlations of the TJC-28 (using θ as well as the sum score) with the sum scores of the established measures of disease activity at baseline (n = 457). Except where indicated, $p \leq 0.01$ (2 tailed).

Measurement of Disease Activity	Correlation with θ TJC-28	Correlation with Sum Score TJC-28
VAS Pain	0.279	0.280
HAQ-ADI	0.405	0.416
VAS GH	0.305	0.302
SJC-28	0.453	0.440
ESR	0.023 (p = 0.626)	0.064 (p = 0.177)

TJC-28: tender joint count for 28 joints; VAS: visual analog scale; HAQ-ADI: Health Assessment Questionnaire-Alternative Disability Index; GH: patient's general health assessment; SJC-28: swollen joint count for 28 joints; ESR: erythrocyte sedimentation rate.

indicates that the θ scores and the sum scores corresponded highly to each other and that attenuation did not pose any serious problems in our study, diminishing the actual advantage of using IRT-based scores instead of sum scores for evaluating the external construct validity of the TJC-28. It also suggests that it is adequate to use sum scores for the calculation of a patient's TJC.

The unequal discrimination parameters give additional support for use of the 2-PL instead of the Rasch model, since those parameters are assumed to be equal in the Rasch model. The parameter results also showed that the smaller joints especially showed high discrimination parameters, indicating that the MCP and PIP joints discriminate better between patients with different degrees of joint tenderness (θ) than do larger joints (shoulder, elbow, wrist, knee; Table 3). This is in line with the clinical experience of healthcare providers treating patients with RA. A point of interest regarding the joint difficulties is that the wrists show the lowest values (Table 3). They also have the highest average score (left wrist: 0.35, right wrist: 0.38), which is consistent with the clinical experience that the wrist is a commonly affected joint in RA[39,40]. This is also reflected in the minor degree of information the wrists provide to the test. However, the wrists do provide some information at the lower levels of disease activity, which can be regarded as a positive property given the limited measurement range of the instrument along the lower range of disease activity.

The results concerning the reflection of left-right symmetry of joint involvement in the joint parameters reflect several studies that emphasize that symmetry of joint involvement characterizes RA[4,5,6,7], providing additional support for the construct validity of the TJC-28. From a strict test perspective it can be argued that this would imply that half of the joints can be removed from the TJC-28. After all, it can point to redundant items, which might be locally dependent and that make the test unnecessarily long. However, removing items might have an effect on the psychometric characteristics of the test by reducing the test information and its corresponding reliability. Moreover, from a clinical perspective it is probably undesirable to remove half of the joints, because a patient's total number of tender joints are being used for individual diagnosis and treatment decisions.

The IRT analyses showed that the TJC-28 is a highly reliable instrument; however, this does not imply that the scale also has high interrater reliability. Interrater bias might still be embedded in the inaccuracies of the measure. It is clear, however, that this type of bias did not pose any serious problems in our study, since problems with interrater reliability have mainly been reported for graded or weighted joint counts[19,41,42], while a nongraded TJC was used in our study.

The accuracy and broadness of the test, given the high discrimination parameters and the range of joint difficulties covered, make accurate measurement of change over time possible. The advantage of using IRT instead of the more traditional approaches is that latent trait values are used instead of sum scores. Even when there are data missing, the latent trait values can still be estimated.

IRT has been successfully applied for the evaluation and improvement of questionnaires of health outcome measures[17]. Since the focus of IRT is at the item level instead of the test level, the contribution of each single joint can be evaluated without knowledge of the other joints in the instrument[24], a feature that is not available in procedures based on CTT methods. Among others, this feature makes it possible to obtain joint counts with lesser joints without major loss of measurement precision[24]. However, IRT has rarely been applied for the evaluation or improvement of clinical measures, such as TJC. One demonstration of the application of IRT in a clinical trial can be found in Glas, *et al*[43]. They successfully applied IRT to tender point counts in fibromyalgia. They showed that tender point counts of patients diagnosed with fibromyalgia had a good fit with IRT models, and that items could be removed without facing a substantial loss of power. Our study extended this application to clinical measures by applying IRT to TJC in patients with early-stage RA. Future studies could investigate whether a modified or shorter TJC will perform equally well or perhaps even better than the TJC-28.

In contrast to CTT, IRT information curves can be obtained when applying IRT to the data. This provides insight into the performance of the total TJC-28 and of the individual joints, and exposes opportunities for scale improvement. The covered range of joint difficulties demonstrated that the TJC-28 mainly functions along the moderate and higher spectrum of disease activity. The test and item information functions also showed that θ is measured with the greatest precision for patients with a higher degree of joint tenderness, especially with joint tenderness in the smaller joints. This spectrum limitation was caused by the low number of painful joints experienced by the sample of patients with early-stage RA. Since the cohort we used represented a large sample size, and since it included

patients from 6 hospitals from different regions in The Netherlands, it is expected that this cohort is representative of the patients with early-stage RA. However, to further examine the measurement precision of the TJC-28 and to make the results more generalizable, future research should expand our study by applying IRT to RA samples with a longer disease duration.

The rationale concerning which joints to include in a joint count has not yet been clearly outlined in the literature. The joints included in the TJC-28 were selected based on pragmatic logistic considerations and clinical experience[20]. Although the TJC-28 appears to be a reliable and valid instrument to assess joint tenderness, it does not include the feet and ankle joints. There have been several discussions about whether the feet and ankles really can be omitted from the instrument[44,45]. It has been argued that the 28-joint count might be useful in clinical trials, but that a more comprehensive joint count that includes the foot joints might be preferable for following the disease progress of patients in daily clinical practice[13,19]. IRT may provide clarity in this discussion, since IRT provides an opportunity to evaluate the contribution of each single foot joint and ankle joint[24]. The joints differ in the degree of information they provide, shown by the inequality in the parameters values. This means the joints contribute unequally to the precision of measurement. By evaluating whether foot and ankle joints provide any significant information, it can be decided whether they truly can safely be omitted from the joint count. Future research should apply IRT to more extensive joint counts, such as the TJC-68, to examine which joints provide important information to the instrument and should be included, and which joints provide limited information and can therefore be omitted from the joint count.

Our study confirmed that the TJC-28 has good internal and acceptable external construct validity for patients with early-stage RA. However, the IRT analyses also pointed to some potential limitations of the instrument — a major problem being its limited measurement range. Since test information was low for lower levels of disease activity, it might be appropriate to modify the TJC-28 to improve its measurement precision and range, for instance by expanding the TJC to joints that provide more information at the lower levels of disease activity. It is recommended that future studies examine both the TJC-28 and more extensive joint indices in RA samples with a longer disease duration to confirm our findings and to explore possibilities for further improvements of the TJC.

ACKNOWLEDGMENT

The authors thank the respondents who participated in our study and the rheumatologists from the Medisch Spectrum Twente (Enschede), Ziekenhuisgroep Twente (Almelo, Hengelo), Isala Klinieken (Zwolle), Universitair Medisch Centrum Groningen (Groningen), UMC St. Radboud (Nijmegen), and TweeSteden Ziekenhuis (Tilburg) for their assistance in patient recruitment and data collection.

REFERENCES

1. Gonzalez A, Maradit Kremers H, Crowson CS, Nicola PJ, Davis JM, Therneau TM, et al. The widening mortality gap between rheumatoid arthritis patients and the general population. Arthritis Rheum 2007;56:3583-7.
2. Tobón GJ, Youinou P, Saraux A. The environment, geo-epidemiology, and autoimmune disease: Rheumatoid arthritis. J Autoimmun 2010;35:10-4.
3. Turkiewicz AM, Moreland LW. Rheumatoid arthritis. In: Bartlett SJ, editor. Clinical care in the rheumatic diseases. Atlanta: Association of Rheumatology Health Professionals; 2006:157-66.
4. Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF, Cooper NS, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31:315-24.
5. Helliwell PS, Hetthen J, Sokoll K, Green M, Marchesoni A, Lubrano E, et al. Joint symmetry in early and late rheumatoid and psoriatic arthritis. Arthritis Rheum 2000;43:865-71.
6. Abramson JH. On the diagnostic criteria of active rheumatoid arthritis. J Chron Dis 1967;20:275-90.
7. Ropes MW, Bennett GA, Cobb S, Jacox R, Jessar RA. Proposed diagnostic criteria for rheumatoid arthritis. Ann Rheum Dis 1957;16:118-25.
8. Aletaha D, Smolen JS. Remission of rheumatoid arthritis: Should we care about definitions? Clin Exp Rheumatol 2006;24(6 Suppl 43):S45-51.
9. Scott DL, Houssien DA. Joint assessment in rheumatoid arthritis. Br J Rheumatol 1996;35:14-8.
10. Pala O, Cavaliere LF. Joint counts. In: Bartlett SJ, editor. Clinical care in the rheumatic diseases. Atlanta: Association of Rheumatology Health Professionals; 2006:39-41.
11. Felson DT, Anderson JJ, Boers M, Bombardier C, Chernoff M, Fried B, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. Arthritis Rheum 1993; 36:729-40.
12. Pinals RS, Masi AT, Larsen RA. Preliminary criteria for clinical remission in rheumatoid arthritis. Arthritis Rheum 1981;24:1308-15.
13. Smolen JS, Breedveld FC, Eberl G, Jones I, Leeming M, Wylie GL, et al. Validity and reliability of the twenty-eight-joint count for the assessment of rheumatoid arthritis activity. Arthritis Rheum 1995;38:38-43.
14. Prevoo MLL, van Riel PLCM, van 't Hof MA, van Rijswijk MH, van Leeuwen MA, Kuper HH, et al. Validity and reliability of joint indices. A longitudinal study in patients with recent onset rheumatoid arthritis. Br J Rheumatol 1993;32:589-94.
15. Prevoo MLL, van 't Hof MA, Kuper HH, van Leeuwen MA, van de Putte LBA, van Riel PLCM. Modified disease activity scores that include twenty-eight-joint counts. Arthritis Rheum 1995; 38:44-8.
16. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Rheum 2007;57:1358-62.
17. Hays RD, Morales LS, Reise SP. Item response theory and health outcome measurement in the 21st century. Med Care 2000;38(9 Suppl):II28-42.
18. Vermeer M, Kuper H, Hoekstra M, Bernelot Moens H, van Riel P, van de Laar M. Remission in daily clinical practice: excellent results after one year of tight control in very early rheumatoid arthritis, results of the DREAM remission induction cohort. Ann Rheum Dis 2010;69:506.
19. Van Riel PLCM, Fransen J, Scott DL. EULAR handbook of clinical assessments in rheumatoid arthritis. Alphen aan den Rijn: van

Zuiden Communications; 2004.
20. Fuchs HA, Brooks RH, Callahan LF, Pincus T. A simplified twenty-eight-joint quantitative articular index in rheumatoid arthritis. Arthritis Rheum 1989;32:531-7.
21. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137-45.
22. Bruce B, Fries JF. The Stanford Health Assessment Questionnaire: a review of its history, issues, progress, and documentation. J Rheumatol 2003;30:167-78.
23. Van den Berg SM, Glas CAW, Boomsma DI. Variance decomposition using an IRT measurement model. Behav Genet 2007;37:604-16.
24. Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of item response theory. Newbury Park, CA: Sage Publications; 1991.
25. Baker FB. The basics of item response theory. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation; 2001.
26. Glas CAW. Modification indices for the 2-pl and the nominal response model. Psychometrika 1999;64:273-94.
27. Te Marvelde JM, Glas CAW, van Landeghem G, van Damme J. Application of multidimensional item response theory models to longitudinal data. Educ Psychol Meas 2006;66:5-34.
28. van Groen MM, ten Klooster PM, Taal E, van de Laar MAFJ, Glas CA. Application of the Health Assessment Questionnaire disability index to various rheumatic diseases. Qual Life Res 2010; 19:1255-63.
29. Scheerens J, Glas CAW, Thomas SM. Educational evaluation, assessment, and monitoring. A systematic approach. Lisse, Netherlands: Swets & Zeitlinger; 2003.
30. Glas CAW, Verhelst ND. Testing the Rasch model. In: Fischer GH, Molenaar IW, editors. Rasch models: their foundations, recent developments and applications. New York: Springer; 1995:69-96.
31. Reeve BB, Fayers P. Applying item response theory modelling for evaluating questionnaire item and scale properties. In: Fayers PM, Hays RD, editors. Assessing quality of life in clinical trials: methods and practice. Oxford: Oxford University Press; 2005:55-73.
32. Van Riel PLCM. Provisional guidelines for measuring disease activity in clinical trials on rheumatoid arthritis. Br J Rheumatol 1992;31:793-4.
33. Browne JP, van der Meulen JH, Lewsey JD, Lamping DL, Black N. Mathematical coupling may account for the association between baseline severity and minimally important difference values. J Clin Epidemiol 2010;63:865-74.
34. El Miedany Y, Youssef SS, El Gaafary M. Short-term outcome after anti-tumor necrosis factor-alpha therapy in rheumatoid arthritis: Do we need to revise our assessment criteria? J Rheumatol 2006;33:490-6.
35. Sokka T. Assessment of pain in rheumatic diseases. Clin Exp Rheumatol 2005;23(5 Suppl 39):S77-84.
36. Plant MJ, O'Sullivan MM, Lewis PA, Camilleri JP, Coles EC, Jessop JD. What factors influence functional ability in patients with rheumatoid arthritis. Do they alter over time? Rheumatology 2005;44:1181-5.
37. Van Leeuwen MA, van der Heijde DMFM, van Rijswijk MH, Houtman PM, van Riel PLCM, van de Putte LBA, et al. Interrelationship of outcome measures and process variables in early rheumatoid arthritis. A comparison of radiologic damage, physical disability, joint counts, and acute phase reactants. J Rheumatol 1994;21:425-9.
38. Dwyer KA, Coty MB, Smith CA, Dulemba S, Wallston KA. A comparison of two methods of assessing disease activity in the joints. Nurs Res 2001;50:214-21.
39. Filippucci E, Iagnocco A, Salaffi F, Cerioni A, Valesini G, Grassi W. Power Doppler sonography monitoring of synovial perfusion at the wrist joints in patients with rheumatoid arthritis treated with adalimumab. Ann Rheum Dis 2006;65:1433-7.
40. Baan H, Hoekstra M, Veehof M, van de Laar M. Ultrasound findings in rheumatoid wrist arthritis highly correlate with function. Disabil Rehabil 2011;33:729-33.
41. Thompson PW, Hart LE, Goldsmith CH, Spector TD, Bell MJ, Ramsden MF. Comparison of four articular indices for use in clinical trials in rheumatoid arthritis: Patient, order and observer variation. J Rheumatol 1991;18:661-5.
42. Hart LE, Tugwell P, Buchanan WW, Norman GR, Grace EM, Southwell D. Grading of tenderness as a source of interrater error in the Ritchie articular index. J Rheumatol 1985;12:716-7.
43. Glas CAW, Geerlings H, van de Laar MAFJ, Taal E. Analysis of longitudinal randomized clinical trials using item response models. Contemp Clin Trials 2009;30:158-70.
44. Landewé R, van der Heijde D, van der Linden S, Boers M. Twenty-eight-joint counts invalidate the DAS28 remission definition owing to the omission of the lower extremity joints: A comparison with the original DAS remission. Ann Rheum Dis 2006;65:637-41.
45. Van der Leeden M, Steultjens MPM, Ursum J, Dahmen R, Roorda LD, van Schaardenburg D, et al. Prevalence and course of forefoot impairments and walking disability in the first eight years of rheumatoid arthritis. Arthritis Rheum 2008;59:1596-602.