Qual Life Res (2009) 18:347–358
DOI 10.1007/s11136-009-9449-3

The 6-item CTS symptoms scale: a brief outcomes measure for carpal tunnel syndrome

Isam Atroshi · Per-Erik Lyrén · Christina Gummesson

Accepted: 27 January 2009 / Published online: 20 February 2009

Abstract

Objective To develop a psychometrically appropriate brief symptoms measure of carpal tunnel syndrome (CTS).
Methods Preoperative CTS 11-item symptom severity and 8-item functional status scales from 693 patients (71% women) with CTS were subjected to exploratory factor analysis and item response theory (IRT) analysis yielding a revised CTS symptoms scale. A validation sample of 213 patients (68% women) with CTS completed the 11-item disabilities of the arm, shoulder and hand (*Quick*DASH), and the revised symptoms scale and 116 patients also completed the original CTS symptom severity scale (median interval 11 days).
Results Of the 11 CTS symptom severity scale items, 2 items that on factor analysis associated with the functional status items were removed. After IRT recalibrations of the remaining symptom severity scale items, 2 non-fitting items were removed and 2 items were merged creating the 6-item CTS symptoms scale. Factor analysis showed one dominant factor explaining 58% of the variance. Reliability was high (Cronbach alpha = 0.86; IRT person separation reliability = 0.88). No item displayed significant differential item functioning. The 6-item CTS symptoms scale showed strong correlation with the *Quick*DASH ($r = 0.70$) and agreement with the original symptom severity scale (ICC = 0.80).
Conclusion The 6-item CTS symptoms scale has good reliability and validity and can be used to measure symptom severity and treatment outcome in CTS.

Keywords Item response theory · Patient-reported outcomes · Symptom severity scale

Abbreviations

CTS	Carpal tunnel syndrome
DASH	Disabilities of the arm, shoulder and hand
EFA	Exploratory factor analysis
IRT	Item response theory
PCM	Partial credit model

Introduction

The impact of carpal tunnel syndrome (CTS) on patients involves mainly and often only symptoms and disability, while motor and sensory impairments are present in less than half the patients even in surgical populations [1–4]. Therefore, symptoms and disability should be the primary outcomes to be measured in clinical studies evaluating treatment effect in CTS, and this is best achieved with validated patient-reported outcomes measures.

The CTS questionnaire, consisting of the symptom severity scale and the functional status scale [5], has been

I. Atroshi
Department of Clinical Sciences, Lund University, Lund, Sweden

I. Atroshi (✉)
Department of Orthopedics, Hässleholm and Kristianstad Hospitals, 28125 Hässleholm, Sweden
e-mail: Isam.Atroshi@skane.se

P.-E. Lyrén
Department of Educational Measurement, Umeå University, 90187 Umea, Sweden
e-mail: Per-Erik.Lyren@edmeas.umu.se

C. Gummesson
Department of Health Sciences, Division of Physiotherapy, Lund University, 22100 Lund, Sweden
e-mail: Christina.Gummesson@med.lu.se

extensively used in clinical studies in several countries [6–8] and the scales have demonstrated good validity, reliability, and interpretability [9]. However, these scales were developed without examining their latent structure using, for example, factor analysis. To our knowledge, the latent structure of the symptom severity and functional status scales has subsequently been examined in only two translated versions of the scales [10, 11]. Moreover, all previous studies that assessed the CTS scales used classical test methodology and involved relatively small populations. Measurement methodology based on item response theory (IRT) is being increasingly used in outcomes research because of its advantages over classical test theory [12]. In IRT models, person and item parameters are on the same scale, which enables person-independent item calibration and item-independent estimation of the latent trait. The latent trait (in this case, symptoms of CTS and hand-related disability) is a continuum on which both items and persons are located [13, 14]. In addition, IRT is useful in evaluating whether items on a scale exhibit differential item functioning (DIF), which means that certain items perform differently in different subgroups such as women and men.

Although the CTS symptom severity scale (11 items) is relatively brief, a shorter symptoms scale that maintains good measurement properties would be more efficient when used in clinical practice and research. In clinical research, several questionnaires may be needed to collect all the information sought and in clinical practice even a small gain in administration time for a common condition such as CTS would be valuable. A brief CTS symptoms scale can increase patient acceptance and improve response rates. In addition, a brief scale would be less frequently associated with missing item responses that may adversely affect validity of the estimates. Although IRT can be used to manage missing item responses, because when the IRT model fits the data estimation of the latent trait is item-independent, enabling trait estimates to be derived from any set of items that are from the same continuum, this is not always possible in clinical practice. Consequently, using the shortest possible measure that adequately provides the information needed would be important. IRT-based methodology is being increasingly used for shortening outcomes questionnaires [15, 16]. By examining each item's properties and location and the distribution of the items on the scale, IRT-based analysis is useful both in optimal scale shortening and in assessing the performance of the reduced scale [17].

The purpose of this study was to develop a psychometrically appropriate brief symptoms measure of CTS based on the CTS symptom severity scale, using exploratory factor analysis and IRT methodology and to assess its reliability and construct validity.

Methods

Samples

Development sample

During a 5-year period (2001 through 2005), the original CTS questionnaire was administered to all patients undergoing carpal tunnel release at one orthopaedic department. In each patient, the diagnosis of CTS was established by the examining surgeon based on history and physical examination and, when judged necessary, nerve conduction tests. The patients completed the symptom severity and functional status scales immediately before surgery. The inclusion criteria were patient age 18 years or older, diagnosis of primary idiopathic CTS, and planned carpal tunnel release with no additional procedures. Only one assessment per patient was used; in patients who had bilateral surgery during the study period, only the questionnaire related to the hand that was operated on first was used.

Preoperative questionnaires were available from 693 patients; 494 women with mean age of 49 (SD 14) years and 199 men with mean age of 49 (SD 13) years. Data from this sample were used in the analyses of the original symptom severity and functional status scales and in revising the CTS symptom severity scale.

Validation sample

For validation of the brief CTS symptoms scale, data from another sample enrolled from January 2007 through June 2008 were used. The validation sample comprised 213 patients diagnosed with primary idiopathic CTS; 145 women with mean age of 52 (SD 17) years and 68 men with mean age of 55 (SD 16) years. Of these 213 patients, 187 were scheduled for carpal tunnel release surgery and 26 were receiving non-operative treatment. These patients completed the revised symptoms scale and the 11-item disabilities of the arm, shoulder and hand (*Quick*DASH).

The last 130 consecutive patients in the validation sample were asked to also complete the original 11-item symptom severity scale in addition to the revised symptoms scale in random order on two separate occasions. The questionnaires together with the *Quick*DASH were sent to the patients by mail. A reminder was sent to those who did not respond within 2 weeks. Of the 130 patients, 116 (89%) returned completed questionnaires; 74 women with mean age of 52 (SD 17) years and 42 men with mean age of 55 (SD 18) years. The mean time interval between completing the two questionnaires was 13 days (median 11, range 2–40). In addition, a test–retest analysis was done among 24 patients (15 women) who completed the revised symptoms scale twice with a mean interval of 14 days (median 10, range 1–35).

Scales

CTS symptom severity and functional status scales

The CTS symptom severity scale consists of 11 items that inquire about severity and frequency of symptoms (night and daytime numbness, tingling, pain, weakness) [5]. The functional status scale consists of eight items that inquire about difficulties in performing specified daily activities. Each item has five response options scored 1 (no symptom or no difficulty in performing the activity) through 5 (most severe symptom or inability to perform the activity). The symptom severity and functional status score is the mean of all answered items in each scale; higher score indicates worse symptoms or disability.

QuickDASH

The *Quick*DASH is a validated 11-item measure of upper-extremity related disability and has been used in patients with CTS [18, 19]. The *Quick*DASH score may range from 0 (no disability) to 100 (most severe disability).

Analyses

Latent structure

The CTS symptom severity and functional status scales were examined using factor analysis. Because the underlying factor structure of the two CTS scales has not been established, no assumptions exist regarding the type or number of factors in each scale and the degree of validity of each possible factor as opposed to the validity of the whole scale which has been previously demonstrated. Because the objective was to derive a brief CTS symptoms scale that measured the essential symptoms it was not expected that the short scale would necessarily maintain all the possible factors or content domains in the original CTS symptom severity scale. We therefore performed an exploratory factor analysis (EFA) of both scales combined to identify the factor structure.

Traditional EFA assumes data are continuous, normally distributed, and without missing values. Of the 693 participants (development sample), 138 (20%) had missing values on at least one item. To examine whether these missing values were missing completely at random (MCAR), Little's MCAR test [20] was performed using SPSS. The test showed that the missing values were missing completely at random ($\chi^2 = 1045.264$, $df = 1030$, $P = 0.36$) and, therefore, the 138 participants with missing values were removed from the EFA sample. The EFA was performed using the computer program FACTOR [21] because of its flexibility with regard to choice of matrix of association, procedures for determining the number of factors (i.e. retention criteria), and rotation techniques. Because the data are based on items with ordered categories they cannot be assumed to be truly continuous or normally distributed. However, data with at least five categories can be treated as essentially continuous [22]. Further, to address this issue, it has been proposed that the correlation matrix based on polychoric correlations rather than Pearson correlations should be analyzed [23]; this is done routinely in FACTOR. However, an initial analysis of the scales showed that the absolute values of the univariate skewness and kurtosis were below 1.0 for 12 of the 19 items and 1.22 was the largest absolute value for all items, which indicates that it would be preferable to analyze the Pearson correlation matrix rather than the polychoric correlation matrix [21]. Analysis of multivariate normality indicated that the data were non-normal (Mardia's coefficient for multivariate kurtosis was 433.13). Using maximum likelihood (ML) estimation as the extraction method assumes multivariate normality. While ML estimation with non-normal (kurtotic) data may produce biased chi-square based fit values and standard error estimates, it still tends to yield stable factor coefficients [24–26]. We therefore concluded that ML estimation was appropriate to use in this study.

The retention rules used were Kaiser's criterion of eigenvalue greater than 1.0, Velicer's minimal average partial (MAP) test, and Horn's parallel analysis [27]. ML estimation was used as the extraction method and the oblique direct oblimin rotation technique was applied because of the possibility that the factors are correlated. Results of the factor analysis were also used in assessing IRT model assumption; in order to regard the unidimensionality assumption as essentially met, the first factor should be dominant and account for more than 20% of the variability, and the first eigenvalue should be at least four times the second eigenvalue [28].

Item analysis

The symptom severity and functional status scales were examined using IRT-based analysis including DIF analysis concerning gender.

The basic idea underlying IRT is that a person's response to a particular item depends on that person's level of the latent trait of interest and the difficulty of that item, and that these two variables can be placed on the same scale [13]. Item response models are models of the probability of a respondent's score on an item, conditional on that respondent's trait level. IRT focuses on the relationship between each individual item and the latent trait and estimates the latent trait by estimating the probabilities that individuals with a certain level of the trait would respond to the items in a certain manner [13]. IRT models calculate standard errors of measurements according to individual trait estimates rather than a single error for all persons in a sample. One of the simplest item response models is the dichotomous Rasch model [29], also known as the 1-parameter (logistic) model. One of the commonly used models for analyzing polytomous items is the Partial Credit Model (PCM) [30], which is an extension of the dichotomous Rasch model.

The data from the development sample were subjected to IRT and DIF analyses (item calibration and model fit) using ConQuest item modeling software [31]. The PCM was used because in a Rasch model the order of the items in terms of item location on the scale is the same for all participants, which is not the case when using non-Rasch models [14]. This is a desirable feature when comparing different participants' performance on the scale; for instance, examining the scale for DIF would involve examining differences in item locations only rather than differences between item response functions at several locations across the scale. Also, when using the PCM (or other Rasch models), the total score is an adequate statistic for estimating the person parameter, which means that using the PCM requires smaller sample sizes than does the non-Rasch IRT models and that it is easy to calculate total scores for participants [13].

ConQuest allows for DIF to be incorporated in the item response modeling, which means that ability estimates, calibration data, and DIF values are obtained in a single analysis. The DIF values provided by ConQuest are the differences between the item location estimates for the specific group and the item location estimates for all respondents. When studying DIF values, two aspects are considered: significance (non-significant values indicate no DIF is present) and magnitude (significant DIF value is of practical importance only if its magnitude reaches a certain level). In examining for significance, the DIF value is considered in relation to its standard error; if the ratio of the DIF value to its standard error is 2, the significance level is approximately 0.05. Consequently, a ratio exceeding 2 indicates significant DIF. Further, the proposed criteria for the magnitude of DIF suggest that a logit difference (i.e. difference in item location) below 0.43 is “negligible”, a value between 0.43 and 0.64 is “intermediate”, and a value greater than 0.64 is “large” [14].

Model fit for the IRT models were assessed with the weighted mean squares (MNSQs). The unweighted MNSQs are the squared standardized residuals averaged over persons, and the weighted MNSQs are squared residuals weighted so that responses in which the person and the item are far from each other on the latent trait scale have less influence on the magnitude of the fit statistic [32]. The weighted MNSQs have an expected value of 1.0, with values within a range from 0.75 (=3/4) and an upper bound of 1.33 (=4/3) indicating reasonable fit [14].

Item selection

For the purpose of deriving a brief symptoms scale from the original symptom severity scale, item selection was based on a decision sequence similar to that described by Cole et al. [33], which involved item fit and amount of severity measured, with the addition of three steps. An initial step was based on the EFA of both scales to check for overlap in the measured constructs and two concluding steps were a DIF analysis and a professional review of the symptoms scale to check for incongruities in item content and format. Consequently, the first step of our decision sequence was to examine the factor structure of both the functional status and symptom severity scales and delete symptom severity scale items that associate with the function rather than the symptom factors. The second step was to check item fit and delete items with poor fit statistics. This step is iterative as poor fitting items are removed one by one, because after one item has been removed and the remaining items are recalibrated the fit statistics will change. The procedure of item removal and recalibration is repeated until all items have acceptable fit. The third step was to look for overlap in the amount of symptom severity measured. If two or more items had the same item location estimate after taking into account the standard errors of the estimates, the item with the best fit was retained [33]. The fourth step was to check for DIF; items with significant non-negligible DIF were considered for removal after examining possible causes (bias) of the DIF related to content and item wording [34]. The final step was to check for incongruity in item content and format.

Reliability

Internal consistency was assessed with the Cronbach's alpha coefficient (α) and the person separation reliability was derived from the IRT analysis [14]. Test–retest reliability was assessed by calculating the intraclass correlation coefficient (absolute agreement) and the mean difference and 95% confidence interval (CI) between the test and retest scores.

Validity

The brief symptoms scale derived from revising the original scale was assessed for validity [35]. In the entire validation sample, the correlations between the revised scale and the *Quick*DASH score (convergent validity) was assessed with the Pearson's correlation coefficient. In the validation subsample that responded to the original and the revised scales, the mean difference between the scores for the original and revised scales and 95% CI were calculated and the agreement between the scores was assessed with the intraclass correlation coefficient.

In all statistical tests a P value of 0.05 was used to indicate statistical significance.

The study was approved by the local Ethics Committee.

Results

Development results

Original CTS symptom severity and functional status scales

Latent structure All functional status scale items and symptom severity scale items were entered into a factor analytic model and the Pearson correlation matrix was examined. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was 0.91 indicating that the items were suitable for factor analysis. The retention rules were not consistent with regard to the number of factors that could be extracted; employing the criterion of eigenvalue greater than 1.0 indicated three factors, the MAP test indicated two factors, and the parallel analysis indicated three factors. However, because two of the three rules were in agreement, three factors were extracted. The analysis showed the presence of a dominant factor that explained 43% of the variance and seemed to be related to function (Table 1). All items in the functional status scale as well as two items from the symptom severity scale, item 7 (weakness) and item 11 (gripping small objects), associated with the first factor (using a criterion for association of >0.4). The symptom severity scale items 1, 2, 6, 8, 9, and 10 (night-time symptoms and numbness/tingling) associated with the second factor, and items 3, 4, and 5 (daytime pain symptoms) associated with the third factor.

Reliability The internal consistency of the symptom severity scale was high ($\alpha = 0.86$). The mean item-total correlation for the symptom severity scale was 0.55. Three items (4, 5 and 11) had item-total correlations <0.5 (Table 2).

Item analysis In the IRT analysis of the symptom severity scale, item locations, associated standard errors, and fit statistics suggested that the PCM seemed to fit most items (Table 2); however, item 5 had a weighted MNSQ outside the desired range. The person separation reliability was 0.87 indicating that the scale was able to efficiently

Table 1 Pattern coefficients (*P*) and structure coefficients (*S*) for the CTS functional status scale and symptom severity scale items

Item[a]		Factor I		Factor II		Factor III	
		P	S	P	S	P	S
F1	Writing	**0.552**	0.614	0.101	0.372	0.029	0.331
F2	Buttoning clothes	**0.739**	0.701	−0.061	0.283	−0.019	0.322
F3	Holding a book	**0.609**	0.672	0.246	0.499	−0.111	0.266
F4	Gripping the telephone	**0.562**	0.608	0.232	0.456	−0.131	0.219
F5	Opening jars	**0.769**	0.747	−0.102	0.279	0.054	0.396
F6	Doing household chore	**0.733**	0.777	0.02	0.39	0.071	0.434
F7	Carrying grocery bag	**0.686**	0.735	0.01	0.364	0.089	0.427
F8	Bathing and dressing	**0.716**	0.692	−0.066	0.278	0.016	0.343
S1	Pain—night	0.014	0.454	**0.625**	0.727	0.295	0.505
S2	Pain—wakening, frequency	−0.051	0.403	**0.644**	0.719	0.304	0.488
S3	Pain—daytime, severity	0.194	0.588	0.128	0.442	**0.682**	0.818
S4	Pain—daytime, frequency	0.049	0.479	0.038	0.336	**0.846**	0.882
S5	Pain—daytime, duration	−0.002	0.365	−0.061	0.201	**0.812**	0.791
S6	Numbness—severity	0.303	0.509	**0.508**	0.628	−0.071	0.242
S7	Weakness—severity	**0.639**	0.693	−0.018	0.327	0.128	0.434
S8	Tingling—severity	0.281	0.479	**0.466**	0.583	−0.047	0.241
S9	Numbness/tingling—night, severity	0.013	0.384	**0.881**	0.857	−0.094	0.198
S10	Numbness/tingling—wakening, frequency	−0.016	0.371	**0.789**	0.79	0.027	0.275
S11	Gripping small objects	**0.675**	0.711	0.016	0.355	0.059	0.393

For each factor item associations >0.40 are shown in bold

[a] *F* Functional status scale, *S* symptom severity scale

Table 2 Item–total correlation, item location estimates, standard errors, and fit statistics for the original 11-item symptom severity scale in the development sample

Item	ITC	Item location	SE	Unweighted fit				Weighted fit				DIF[a]	
				MNSQ	CI		*t*	MSNQ	CI		*t*	Estimate	SE
S1	0.66	0.16	0.05	0.82	0.89	1.11	−3.6	0.80	0.90	1.10	−4.2	0.01	0.03
S2	0.65	0.26	0.05	0.84	0.89	1.11	−3.1	0.86	0.90	1.10	−2.8	−0.03	0.03
S3	0.65	0.67	0.09	0.79	0.89	1.11	−4.2	0.79	0.90	1.10	−4.3	0.11	0.04
S4	0.48	−0.50	0.04	1.02	0.89	1.11	0.3	1.06	0.90	1.10	1.1	−0.03	0.03
S5	0.44	−0.35	0.04	1.48	0.89	1.11	7.7	1.34	0.90	1.10	6.1	−0.01	0.03
S6	0.58	−0.56	0.07	0.96	0.89	1.11	−0.8	0.96	0.90	1.10	−0.7	0.05	0.03
S7	0.51	0.26	0.05	1.15	0.89	1.11	2.7	1.15	0.90	1.10	2.8	−0.17	0.03
S8	0.50	−0.12	0.06	1.01	0.89	1.11	0.2	1.01	0.90	1.10	0.1	0.04	0.03
S9	0.57	−0.40	0.05	1.02	0.89	1.11	0.4	1.03	0.90	1.10	0.5	0.03	0.03
S10	0.59	0.01	0.05	1.00	0.89	1.11	0.1	1.00	0.90	1.10	0.1	−0.03	0.03
S11	0.49	0.57[b]		1.08	0.89	1.11	1.4	1.08	0.90	1.10	1.5	0.03[b]	

See Table 1 for item description

ITC Corrected item-total correlation, *SE* standard error, *MNSQ* mean square, *CI* 95% confidence interval, *t* *t*-score, *DIF* differential item functioning

[a] DIF values are deviations from the item location parameters for the group of women, the difference in item location between women and men is twice the DIF value

[b] The item parameter estimate is constrained

discriminate between respondents. The mean (SD) of the latent trait distribution was −0.25 (0.86) and the mean of the standard error of the latent trait estimates was 0.33. A 95% CI for the latent trait estimates for the original 11-item scale would be approximately ±0.66, corresponding to ±77% of one SD.

DIF analysis of the symptom severity scale (Table 2) showed statistically significant DIF in item location for item 3 (higher for women) and item 7 (higher for men), and analysis of the functional status scale showed significant differences in item location for items 1 and 2 (higher for women) and items 5 and 7 (higher for men). However, because sample size has a strong influence on significance tests, the magnitude of the differences in item locations was also considered. The symptom severity scale item 7 demonstrated the largest DIF with a difference in item location of 0.33 (higher for men), which is considered a negligible DIF.

Item selection As a result of the first step in the item selection procedure items 7 and 11 were removed from the scale because they were clearly more associated with the functional status scale than with the symptom severity scale items. After the first step, the remaining items were recalibrated (Table 3). The second step began with the removal of item 5, because the item's fit statistic (a weighted MNSQ of 1.40) was outside the range for acceptable fit. After another recalibration item 4 was similarly removed, because of its weighted MNSQ of 1.49. Following recalibration of the remaining items, all items had acceptable fit statistics. In addition, there was no overlap in item location estimates among the seven remaining items and all had negligible DIF. Thus, four items (items 4, 5, 7, and 11) were removed before examination for item incongruity.

The final step in the item selection procedure revealed an incongruity. In the original scale, item 9 inquires about severity of "numbness or tingling at night" and item 10 inquires about frequency of night wakening because of "numbness or tingling", whereas two separate items inquire about severity of "numbness" (item 6) and severity of "tingling" (item 8). Because of this incongruity and supported by high inter-item correlation (0.71) between items 6 and 8, these two items were merged into one item that inquires about severity of "numbness or tingling during daytime".

To improve the appearance of the questionnaire and facilitate easier and quicker responding to the items we changed the order of the response choices from longitudinal to transverse, which meant that, instead of the original questionnaire's separate stems for each of the 11 items, only two stems were required for the 6 items in the revised scale (Appendix).

Validation results

6-item CTS symptoms scale

Latent structure The factor analysis of the 6-item scale among the 213 patients who responded to the revised scale and had no missing data showed one dominant factor that

Table 3 Item locations, their respective standard error, and weighted mean squares after three item removal sequences and subsequent recalibrations of the original 11-item symptom severity scale in the development sample

Item	Recalibration I			Recalibration II			Recalibration III			DIF[a]	
	Item location	SE	Weighted MNSQ	Item location	SE	Weighted MNSQ	Item location	SE	Weighted MNSQ	Estimate	SE
S1	0.26	0.05	0.83	0.24	0.05	0.83	0.19	0.05	0.83	−0.01	0.03
S2	0.37	0.05	0.87	0.36	0.05	0.89	0.33	0.05	0.90	−0.06	0.03
S3	0.80	0.09	0.84	0.81	0.09	1.02	0.80	0.09	1.25	0.08	0.04
S4	−0.43	0.04	1.11	−0.49	0.05	1.49	–	–	–	0.19	0.19
S5	−0.27	0.04	1.40	–	–	–	–	–	–	0.19	0.19
S6	−0.49	0.07	1.05	−0.56	0.07	1.00	−0.68	0.07	1.04	0.03	0.03
S8	−0.02	0.06	1.08	−0.06	0.06	1.07	−0.13	0.06	1.12	0.02	0.03
S9	−0.33	0.05	1.03	−0.39	0.05	0.90	−0.51	0.05	0.88	0.00	0.03
S10	0.10[b]		1.01	0.07[b]		0.93	0.01[b]		0.94	−0.06[b]	

SE Standard error, *MNSQ* mean square, *DIF* differential item functioning

[a] See footnote in Table 2

[b] The item parameter estimate is constrained

explained 58% of the variance, with all 6 items being associated with that factor (pattern/structure coefficients ranging from 0.69 to 0.84).

Reliability Cronbach's alpha for the six items was 0.86 and the item-total correlations ranged from 0.56 to 0.74 (Table 4). Test–retest reliability was high; ICC was 0.95 (95% CI 0.90–0.98) and the mean 6-item score at the two occasions was 3.28 (SD 0.7) and 3.25 (SD 0.7), respectively (mean difference, 0.03; 95% CI, −0.07 to 0.12).

Item analysis Item locations, their associated standard errors, and fit statistics for the 6-item CTS symptoms scale showed that the PCM fit the scale well, with all the items within the desired weighted MNSQ bounds (Table 4). The items and respondents appeared to match fairly well on the latent trait (Figs. 1, 2). The person separation reliability was 0.88 indicating that the 6-item scale performs similarly well as the 11-item symptom severity scale in separating respondents on the latent trait continuum. The mean (SD) of the latent trait distribution was 0.33 (1.38) and the mean of the standard error of the latent trait estimates was 0.56. The standard error was larger than that for the original scale but the 95% CI is approximately ±81% of one SD, which is similar to that for the original 11-item scale. No item displayed statistically significant DIF.

Validity In the entire validation sample the mean 6-item CTS symptoms score was 3.2 (SD 0.7) and the mean *Quick*DASH score was 51.6 (SD 20), with the two scale scores correlating strongly (Pearson correlation coefficient = 0.70). In the validation subcohort of patients who completed both the original and revised scales, strong

Table 4 Item–total correlation, item location estimates, standard errors, and fit statistics for the 6-item CTS symptoms scale in the validation sample

Item	ITC	Item location	SE	Unweighted fit				Weighted fit				DIF[a]	
				MNSQ	CI		*t*	MSNQ	CI		*t*	Estimate	SE
1 (S1)	0.74	0.34	0.11	0.81	0.81	1.19	−2.0	0.81	0.81	1.19	−2.0	0.03	0.05
2 (S2)	0.60	0.44	0.12	1.03	0.81	1.19	0.4	1.07	0.81	1.19	0.7	−0.02	0.05
3 (S3)	0.62	−0.99	0.15	1.08	0.81	1.19	0.8	1.05	0.81	1.19	0.5	0.00	0.05
4 (S6/S8)	0.56	−0.35	0.13	1.22	0.81	1.19	2.1	1.18	0.81	1.19	1.8	−0.03	0.05
5 (S9)	0.72	0.70	0.11	0.84	0.81	1.19	−1.8	0.83	0.81	1.19	−1.9	0.05	0.05
6 (S10)	0.62	−0.15[b]		1.02	0.81	1.19	0.2	1.01	0.81	1.19	0.1	−0.03[b]	

Item location parameter is the mean of the thresholds for that item (see Fig. 1 for a map of thresholds)

See Table 1 and Appendix for item description; item 4 is the result of merging original items S6 and S8 (see text)

ITC Corrected item-total correlation, *SE* standard error, *MNSQ* mean square, *CI* 95% confidence interval, *t* *t*-score, *DIF* differential item functioning

[a] See footnote in Table 2

[b] The item parameter estimate is constrained

Fig. 1 Map of latent trait distributions and item response parameter estimates for the 6-item CTS symptoms scale in the validation sample. The *X*s indicate the persons (each *X* represents three persons)

```
Latent trait         Items
              |
              |
 5            |
              |
              |
              |
             X|
 4            |
              |
              |
              |
             X|
 3            |
             X|
              |
           XXX|
             X|
 2            |
         XXXXX|
              |
         XXXXX|
        XXXXXX|
 1            |
         XXXXX|
              |5
    XXXXXXXXXX|1  2
         XXXXX|
 0            |
        XXXXXX|6
           XXX|4
              |
          XXXX|
-1        XXXX|3
           XXX|
              |
            XX|
           XXX|
-2            |
             X|
             X|
              |
              |
-3            |
              |
              |
              |
              |
-4            |
              |
              |
              |
              |
-5            |
              |
```

agreement was shown between the scale scores, which was similar to the agreement between the *Quick*DASH scores for the two administration times (Table 5). The 95% confidence interval for the mean difference in scores between the original and revised scale was within 0.15 point on the 1 to 5-point scale and for the *Quick*DASH was within approximately 3 points on the 0 to 100-scale.

Discussion

We have derived a 6-item CTS symptoms scale from the commonly used 11-item symptom severity scale and have shown that the short scale has maintained good measurement properties. The use of the short scale would likely improve patient acceptance and increase the response rate and thus improve the efficiency of outcome measurement in CTS.

We validated the revised scale in a separate sample, as recommended [35], and the IRT analyses showed that the 6-item scale appears to measure the latent trait with an amount of error similar to that of the original 11-item scale. In addition, good agreement was shown between the two scales when administered consecutively to the same population despite a varying time interval. In a test–retest reproducibility of the Dutch version of the CTS symptom

Fig. 2 Map of latent trait distributions and item thresholds for the 6-item CTS symptoms scale in the validation sample. The thresholds are indicated with item number and threshold number (e.g., 2.3 means threshold 3 on item 2). The *X*s indicate the persons (each *X* represents three persons)

```
Latent trait                Generalized-item thresholds
                |
                |
   5            |
                |
                |
                |
               X|
   4            |   2.4
                |
                |            5.4
                |1.4
               X|
   3            |         4.4
               X|               6.4
                |
             XXX|
               X|      3.4
   2            |
           XXXXX|
                |
           XXXXX|            5.3
          XXXXXX|   2.3
   1            |1.3
           XXXXX|         4.3   6.3
                |
      XXXXXXXXXX|
           XXXXX|
   0            |
          XXXXXX|      3.3
             XXX|            5.2
                |
            XXXX|
  -1        XXXX|   2.2
             XXX|1.2
                |
              XX|               6.2
             XXX|         4.2
  -2            |1.1         5.1
               X|      3.2
               X|
                |   2.1
                |               6.1
  -3            |
                |
                |
                |         4.1
                |      3.1
  -4            |
                |
                |
                |
                |
  -5            |
                |
```

Table 5 The scores for the 6-item CTS symptoms scale, the original 11-item symptom severity scale and the *Quick*DASH on the two successive administration times in the validation sample

Scale	*n*	Mean (SD)	Mean difference (95% CI)	*P*	ICC (95% CI)
6-item CTS symptoms	116	3.2 (0.7)			
11-item CTS symptom severity	116	3.3 (0.7)	−0.07 (−0.15–0.02)	0.11	0.80 (0.73–0.86)
*Quick*DASH Time I	113	51.5 (20)			
*Quick*DASH Time II	113	51.0 (19)	0.50 (−1.4–2.3)	0.62	0.87 (0.82–0.91)

Score range for the CTS scale is 1–5 and for the *Quick*DASH is 0–100

ICC Intraclass correlation coefficient, *CI* confidence interval severity scale [36], 84 primary care patients with wrist or hand problems completed the 11-item scale twice within 1–2 weeks (mean 10 days) and the mean score difference was 0.11 and the intraclass correlation coefficient was 0.68. In a previous test–retest reliability (1–3 weeks) of the Swedish version of the 11-item symptom severity scale in 22 patients before carpal tunnel release the score difference was 0.10 [6], which is similar to the test–retest results for the 6-item scale in the present study. These data suggest that the difference between the original 11-item and the 6-item scales measured in our study appears to be similar to the expected test–retest difference for the original scale.

The question of content validity is important when deriving short forms [35]. Numbness and tingling are the fundamental features of CTS and night-time symptoms are characteristic whereas daytime pain maybe an important but is not essential part of the disease. These symptoms are covered in the short-form. Because the purpose was not to create a short-form that maintains all the dimensions of the current questionnaire but rather a brief symptoms scale, the removal of some items altered the content. The two items concerning frequency and duration of daytime pain were removed but pain was still represented on the revised scale. The deleted items may not add essential information considering that CTS is not always a painful condition; a large Dutch population-based study of neuropathic pain conditions showed individuals with CTS had the lowest use of pain medication [37]. The item that inquired about weakness in the original symptom severity scale is not included in the 6-item scale. Weakness is a non-specific symptom and hand strength can be measured directly with other methods. In addition, it is well known that, even when symptoms of CTS had completely resolved after carpal tunnel release surgery, hand strength does not usually return to preoperative level until about 3 months after surgery [38].

In the factor analysis, the item concerning weakness associated clearly more with the factor that represented function than with the symptom factors. Similarly, the item regarding difficulty in gripping small objects is neither a specific nor a dominant feature in CTS. In fact, similar results have been found regarding these items in two previous studies that had subjected the original CTS symptom severity scale to exploratory factor analysis. In a study of the Portuguese version, factor analysis of both scales combined showed that items 7 and 11 associated mainly with the functional status scale items [10]. In a factor analysis of the Japanese version of the symptom severity scale, item 7 had low loadings in a 2-factor model and item 11 had the lowest loading in a 1-factor model [11]. The factor analysis of the 11-item symptom severity scale showed that items 7 and 11 were more associated with functional status items, which questions the inclusion of these two items in a symptoms scale.

The original symptom severity scale used separate stems for each item which required two full pages for the scale's 11 items [5]. The revision of the scale to six items also facilitated the use of one common stem for the first four items and one stem for the last two items, shortening the questionnaire to half a page. It is not known whether the changed layout may influence the response pattern but considering the strong agreement between the responses to the original and revised scales the layout issue does not seem to have influenced the responses.

The functional status scale was not modified in this study because it is relatively short and it has been previously shown [39] not to be specific to CTS. Other measures of hand or arm related disability can be used in combination with the 6-item CTS symptoms scale. One measure is the *Quick*DASH which seems to perform similarly to the CTS functional status scale [9], with the possible advantage of facilitating comparison with other upper extremity conditions [19].

Although the 6-item symptoms scale can be scored using IRT-based methods it can also be scored, similar to the original 11-item symptom severity scale, as the mean of all answered items, which would probably be more common in clinical practice. The original CTS symptom severity scale was published without guidelines regarding how to manage unanswered items. Other scales have employed various rules; the SF-36 requires that at least 50% of the items have been answered [40]. If IRT is used in scoring then missing items could, under certain conditions, be managed based on answered items. However, if the scale is scored as the mean of answered items we believe that it would be appropriate not to accept more than 1 missing item in order to calculate a score. The high agreement between the 6-item symptoms scale and the original 11-item symptom severity scale shown in this study suggests that interpretation of the scores should be similar for the two scales. In previous studies the mean CTS symptom severity score for patients with CTS planned for surgery has ranged from 3.1 to 3.3 [38, 41] and the corresponding mean *Quic*kDASH score has been approximately 50 [19] whereas in patients with CTS undergoing non-operative treatment the mean CTS score was lower by approximately 0.5 point [41]. Following surgery the mean CTS symptom severity score usually improves by 1.2 to 1.6 points [38, 41]. The amount of change in the 6-item CTS symptoms score after various treatments needs to be evaluated in longitudinal studies.

One of the advantages of using a brief symptoms scale is that when several types of outcomes measures need to be combined the overall respondent burden would be reduced. The 6-item CTS symptoms scale can be combined when necessary with the *Quick*DASH, which is mainly a measure of activity limitations, or with utility measures such as the SF-6D or the EQ-5D to measure cost-effectiveness [42]. Although the type of measure used is dictated by the objectives of the study, a measure of CTS symptoms should be the minimum outcome measure when evaluating or comparing the efficacy of treatments. With the 6-item CTS symptoms scale this should be easily accomplished both in clinical and research settings.

Acknowledgment Funding This research was supported by Skane county council's research and development foundation, Hässleholm Hospital, and Umeå University, Sweden.

Appendix

The 6-item CTS symptoms scale

The following questions refer to your symptoms for a typical 24-h period during the past 2 weeks. Mark one answer to each symptom

How severe are the following symptoms in your hand?	None	Mild	Moderate	Severe	Very severe
Pain at night					
Pain during daytime					
Numbness or tingling at night					
Numbness or tingling during daytime					

How often did the following symptoms in your hand wake you up at night?	Never	Once	2 or 3 times	4 or 5 times	More than 5 times
Pain					
Numbness or tingling					

References

1. Katz, J. N., Gelberman, R. H., Wright, E. A., Lew, R. A., & Liang, M. H. (1994). Responsiveness of self-reported and objective measures of disease severity in carpal tunnel syndrome. *Medical Care, 32*, 1127–1133. doi:10.1097/00005650-199411000-00005.
2. Atroshi, I., Gummesson, C., Johnsson, R., & Sprinchorn, A. (1999). Symptoms, disability, and quality of life in patients with carpal tunnel syndrome. *The Journal of Hand Surgery, 24*, 398–404. doi:10.1053/jhsu.1999.0398.
3. Agabegi, S. S., Freiberg, R. A., Plunkett, J. M., & Stern, P. J. (2007). Thumb abduction strength measurement in carpal tunnel syndrome. *The Journal of Hand Surgery, 32*, 859–866. doi:10.1016/j.jhsa.2007.04.007.
4. Mallette, P., Zhao, M., Zurakowski, D., & Ring, D. (2007). Muscle atrophy at diagnosis of carpal and cubital tunnel syndrome. *The Journal of Hand Surgery, 32*, 855–858. doi:10.1016/j.jhsa.2007.03.009.
5. Levine, D. W., Simmons, B. P., Koris, M. J., Daltroy, L. H., Hohl, G. G., Fossel, A. H., et al. (1993). A self-administered questionnaire for the assessment of severity of symptoms and functional status in carpal tunnel syndrome. *The Journal of Bone and Joint Surgery. American Volume, 75*, 1585–1592.
6. Atroshi, I., Johnsson, R., & Sprinchorn, A. (1998). Self-administered outcome instrument in carpal tunnel syndrome: reliability, validity and responsiveness evaluated in 102 patients. *Acta Orthopaedica Scandinavica, 69*, 82–88.
7. Mondelli, M., Reale, F., Sicurelli, F., & Padua, L. (2000). Relationship between the self-administered Boston questionnaire and electrophysiological findings in follow-up of surgically-treated carpal tunnel syndrome. *Journal of Hand Surgery (Edinburgh, Lothian), 25*, 128–134. doi:10.1054/jhsb.2000.0361.
8. Rosales, R. S., Delgado, E. B., & Diez de la Lastra-Bosch, I. (2002). Evaluation of the Spanish version of the DASH and carpal tunnel syndrome health-related quality-of-life instruments: cross-cultural adaptation process and reliability. *The Journal of Hand Surgery, 27*, 334–343. doi:10.1053/jhsu.2002.30059.
9. Leite, J. C., Jerosch-Herold, C., & Song, F. (2006). A systematic review of the psychometric properties of the Boston Carpal Tunnel Questionnaire. *BMC Musculoskeletal Disorders, 7*, 78. doi:10.1186/1471-2474-7-78.
10. de Campos, C. C., Manzano, G. M., Leopoldino, J. F., Nobrega, J. A., Sanudo, A., de Araujo, P. C., et al. (2004). The relationship between symptoms and electrophysiological detected compression of the median nerve at the wrist. *Acta Neurologica Scandinavica, 110*, 398–402. doi:10.1111/j.1600-0404.2004.00332.x.
11. Imaeda, T., Uchiyama, S., Toh, S., Wada, T., Okinaga, S., Sawaizumi, T., et al. (2007). Validation of the Japanese Society for surgery of the hand version of the carpal tunnel syndrome instrument. *Journal of Orthopaedic Science, 12*, 14–21. doi:10.1007/s00776-006-1087-9.
12. Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care, 38*, II28–II42. doi:10.1097/00005650-200009002-00007.
13. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.
14. Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

15. Stroud, M. W., McKnight, P. E., & Jensen, M. P. (2004). Assessment of self-reported physical activity in patients with chronic pain: Development of an abbreviated Roland-Morris disability scale. *Journal of Pain, 5*, 257–263. doi:10.1016/j.jpain.2004.04.002.
16. Petersen, M. A., Groenvold, M., Aaronson, N., Blazeby, J., Brandberg, Y., de Graeff, A., et al. (2006). Item response theory was used to shorten EORTC QLQ-C30 scales for use in palliative care. *Journal of Clinical Epidemiology, 59*, 36–44. doi:10.1016/j.jclinepi.2005.04.010.
17. Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16*(Suppl 1), 5–18. doi:10.1007/s11136-007-9198-0.
18. Beaton, D. E., Wright, J. G., & Katz, J. N. (2005). Development of the QuickDASH: Comparison of three item-reduction approaches. *The Journal of Bone and Joint Surgery. American Volume, 87*, 1038–1046. doi:10.2106/JBJS.D.02060.
19. Gummesson, C., Ward, M. M., & Atroshi, I. (2006). The shortened disabilities of the arm, shoulder and hand questionnaire (QuickDASH): Validity and reliability based on responses within the full-length DASH. *BMC Musculoskeletal Disorders, 7*, 44. doi:10.1186/1471-2474-7-44.
20. Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*, 1198–1202. doi:10.2307/2290157.
21. Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods, 38*, 88–91.
22. Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 269–314). Greenwich, CT: Information Age Publishing.
23. Panter, A. T., Swygert, K. A., Grant, D. W., & Tanaka, J. S. (1997). Factor analytic approaches to personality item-level data. *Journal of Personality Assessment, 68*, 561–589. doi:10.1207/s15327752jpa6803_6.
24. Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *The British Journal of Mathematical and Statistical Psychology, 38*, 171–189.
25. Benson, J., & Fleishman, J. A. (1994). The robustness of maximum likelihood and distribution-free estimators to non-normality in confirmatory factor analysis. *Quality & Quantity, 28*, 117–136. doi:10.1007/BF01102757.
26. Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling: A Multidisciplinary Journal, 7*, 557–595.
27. Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.
28. Hambleton, R. K. (2005). Applications of item response theory to improve health outcomes assessment: Developing item banks, linking instruments, and computer-adaptive testing. In J. Lipscomb, C. C. Gotay, & C. Snyder (Eds.), *Outcomes assessment in cancer* (pp. 445–464). Cambridge: Cambridge University Press.
29. Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research. (Reprinted by University of Chicago Press, 1980).
30. Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174. doi:10.1007/BF02296272.
31. Wu, M. L., Adams, R. J., & Wilson, M. (2007). *ConQuest: Generalized Item Response Modeling Software*. Hawthorn, Australia: Australian Council for Educational Research (ACER).
32. Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.
33. Cole, J. C., Rabin, A. S., Smith, T. L., & Kaufman, A. S. (2004). Development and validation of a Rasch-derived CES-D short form. *Psychological Assessment, 16*, 360–372. doi:10.1037/1040-3590.16.4.360.
34. Shepard, L. A. (1985). Identifying bias in test items. In B. F. Green (Ed.), *New directions in testing and measurement: Issues in testing–Coaching, disclosure, and test bias* (pp. 79–104). San Francisco: Jossey-Bass.
35. Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment, 12*, 102–111. doi:10.1037/1040-3590.12.1.102.
36. Spies-Dorgelo, M. N., Terwee, C. B., Stalman, W. A., & van der Windt, D. A. (2006). Reproducibility and responsiveness of the symptom severity scale and the hand and finger function subscale of the Dutch arthritis impact measurement scales (Dutch-AIMS2-HFF) in primary care patients with wrist or hand problems. *Health and Quality of Life Outcomes, 4*, 87. doi:10.1186/1477-7525-4-87.
37. Dieleman, J. P., Kerklaan, J., Huygen, F. J., Bouma, P. A., & Sturkenboom, M. C. (2008). Incidence rates and treatment of neuropathic pain conditions in the general population. *Pain, 137*, 681–688. doi:10.1016/j.pain.2008.03.002.
38. Atroshi, I., Larsson, G. U., Ornstein, E., Hofer, M., Johnsson, R., & Ranstam, J. (2006). Outcomes of endoscopic surgery compared with open surgery for carpal tunnel syndrome among employed patients: Randomised controlled trial. *British Medical Journal, 332*, 1473–1476. doi:10.1136/bmj.38863.632789.1F.
39. Atroshi, I., Breidenbach, W. C., & McCabe, S. J. (1997). Assessment of the carpal tunnel outcome instrument in patients with nerve-compression symptoms. *The Journal of Hand Surgery, 22A*, 222–227. doi:10.1016/S0363-5023(97)80155-4.
40. Ware, J. E., Snow, K. K., Kosinski, M., & Gandek, B. (1993). *SF-36 health survey manual and interpretation guide*. Boston: New England Medical Center.
41. Katz, J. N., Keller, R. B., Simmons, B. P., Rogers, W. D., Bessette, L., Fossel, A. H., et al. (1998). Maine carpal tunnel study: Outcomes of operative and nonoperative therapy for carpal tunnel syndrome in a community-based cohort. *The Journal of Hand Surgery, 23*, 697–710. doi:10.1016/S0363-5023(98)80058-0.
42. Atroshi, I., Gummesson, C., McCabe, S. J., & Ornstein, E. (2007). The SF-6D health utility index in carpal tunnel syndrome. *Journal of Hand Surgery (Edinburgh, Lothian), 32*, 198–202. doi:10.1016/j.jhsb.2006.11.002.