Physical activity in hard-to-reach physically disabled people

Krops, Leonie

Reliability of the Dutch translation of the RAND 36-item health survey in a post-rehabilitation population

Krops LA, Wolthuizen L, Dijkstra PU, Jaarsma EA, Geertzen JHB, Dekker R.

International Journal of Rehabilitation Research,
2018; 41(2):128-137
doi:10.1097/MRR.0000000000000265

2

Abstract

The aim of this study is to evaluate the reliability of the RAND 36-item Health survey as a measure of health related quality of life in a general Dutch post-rehabilitation population. A total of 752 ex-rehabilitation patients were invited to complete the Dutch RAND 36-item health survey. After 2 weeks, the people who responded to the first questionnaire were asked to complete the same questionnaire again. Internal consistency of the questionnaire was expressed as Cronbach's α. Test–retest reliability was expressed as intraclass correlation coefficient (ICC) and presented in Bland–Altman plots. Internal consistency was found acceptable for all subscales (n=276; Cronbach's α ranged from .81 to .95). Test–retest reliability was found acceptable for research and group comparisons for all subscales (n=184; ICC ranged from .71 to .88). Overall, test–retest reliability of the physical functioning (ICC=.86), pain (ICC=.87), and general health (ICC=.88) subscale was relatively high, and that of health change (ICC=.71) was relatively low. Reliability of the questionnaire did not notably differ between participants who indicated stable health and participants who indicated health change during the past weeks. In conclusion, the Dutch translation of the RAND 36-item health survey is reliable for research and group comparisons in a general post-rehabilitation population. However, the RAND 36-item health survey is not sufficiently reliable for individual comparisons within this population.

Introduction

Health related quality of life (HRQoL) is a typical example of a construct that emerged from the bio psychosocial perspective on health, which is the most common view since the past decades[1]. Because of the need for metrics on performance and processes in healthcare in order to improve the quality and effectiveness of care, HRQoL is frequently used as an outcome measure[46]. A number of assessment instruments have been developed to evaluate HRQoL. A short self-administered health survey, such as the multidimensional RAND 36-item health survey 1.0 (RAND-36), is frequently used in rehabilitation[46]. A search in PubMed resulted in 70 studies using the RAND-36 in groups of patients within the rehabilitation population. The RAND-36 is a profile based questionnaire with different subscales: physical functioning, social functioning, role limitations due to physical health problems, role limitations due to personal or emotional problems, mental health, pain, vitality, and general health perception. In addition, one item assesses health change over the past year[47].

The RAND-36 consists of exactly the same 36 items as the 36-Item short-form health survey (SF-36)[48]. The scoring procedure differs somewhat between RAND-36 and SF-36. Only for two subscales, pain and general health, the RAND-36 results in a slightly higher score compared with the SF-36. The other subscales are exactly the same. Overall, the RAND-36 and SF-36 are extremely highly correlated (.99)[49]. The Dutch translation of the RAND-36 was approved as valid[50].

To our knowledge, the reliability of the Dutch translation of the RAND-36 was investigated in only two studies. In healthy individuals, test–retest reliability (Pearson *r)* ranged from .58 to .82[51]. In a brain injury population, test–retest reliability, expressed in intraclass correlation coefficient (ICC), ranged from .44 to .94[52]. As Pearson *r* cannot detect systematic differences, the use of ICC is preferred[53,54]. Appendix 1 provides an indication for reliability values for both the RAND-36 and the SF-36, as found in a search in PubMed. Although reliability of the SF-36 has been tested in groups of patients with different diagnoses within the rehabilitation population[52,55–63], no studies investigated reliability of the Dutch translation of the RAND-36 in a general rehabilitation population. Because of the increased use of HRQoL as an outcome measure in general rehabilitation populations, reliability assessment is essential for an appropriate interpretation of the results of HRQoL assessment.

The aim of this study is to evaluate the reliability of the Dutch translation of the RAND-36 in a general post-rehabilitation population. The reliability of the RAND-36 and SF-36 was acceptable or good in earlier research on different populations[51,52,55–63].

As the health of post-rehabilitation patients may be less stable compared with healthy individuals, reliability in this study is hypothesised to be lower compared with findings in healthy individuals. However, as found in research on other diagnoses, we hypothesise the RAND-36 to be reliable in a post-rehabilitation population.

Methods

Participants

In this study, a total of 752 post-rehabilitation patients were invited to participate. All of them underwent their rehabilitation programme in the Center for Rehabilitation of the University Medical Center Groningen. We invited post-rehabilitation patients instead of clinical patients because health of clinical patients is assumed to be instable owing to treatment and disease progression, which may bias findings on test–retest reliability. Inclusion criteria were diagnoses belonging to the specialism of rehabilitation medicine (for instance, neurological problems, amputation, chronic pain), age of 18 years or older, and sufficient capacity to complete the questionnaire. Excluded were patients with a diagnosis of orthopaedic origin, as they were treated mostly mono-disciplinary, and patients who received cardiac or pulmonary rehabilitation, as they were treated in a different treatment framework. Of all post-rehabilitation patients who were treated between January 2012 and December 2014, and met the above-mentioned criteria, 752 patients were selected, based on a random selection of 94 patients within each of the eight diagnosis categories (amputation, cerebrovascular accident, other brain injuries, chronic pain, multiple sclerosis, spinal cord injury, other neurological disability, and other diagnoses). The test–retest reliability population consisted of all patients who returned the first questionnaire within 4 weeks (n=261).

Questionnaire

In this study, the Dutch translation of the RAND-36 was used[50]. Values for each of the nine subscales range from 0 to 100; higher scores represent a better HRQoL[50]. The RAND-36 is a valid and freely available questionnaire[50]. To determine the stability of the construct under measurement, a question on the experienced health change during the past 4 weeks was added. Participants were asked if their health was much better, a bit better, the same, a bit worse, or much worse compared with 4 weeks ago.

Procedure

The patient's names, addresses, diagnosis, and date of last treatment were retrieved from the database of the Center for Rehabilitation of the University Medical Center Groningen. The questionnaire, including a cover letter, was sent to the potential participants by post. Participants were asked to complete and return the paper questionnaire. After 2 weeks, all participants who returned the first questionnaire were invited to complete exactly the same questionnaire again[64]. Within 2–4 weeks after sending the first questionnaire, a second questionnaire was sent on the day that the first questionnaire was received back. No retests were sent later than 4 weeks after sending the first questionnaire, in order to restrict the interval between the test and retest. All participants signed informed consent, after being informed that participation was voluntary, and that the data would be processed anonymously (except for matching test and retest). The Medical Ethical Committee of the University Medical Center Groningen the Netherlands confirmed that ethical approval was not required for this study (METc 2012.450).

Data analysis

Differences between participants and nonparticipants were analysed using independent-samples *t*-tests (age and follow-up) and x^2-tests (diagnoses and sex). In the entire tested population, the internal consistency (Cronbach's α and the mean inter-item correlation) of the different subscales was analysed. Moreover, floor and ceiling effects were calculated as the percentage of participants who scored the lowest or highest possible score in the entire tested population. Proportions of more than 15% scoring, respectively, the lowest or highest possible score were indicated as substantial floor or ceiling effects[65,66]. In the population that completed the questionnaire twice, a Wilcoxon signed-ranks test was performed to investigate systematic differences between the pretest and post-test because of non-normality of the data. Test–retest reliability was analysed by calculating ICC (two-way random, absolute agreement), Pearson *r* between the test and retest to compare our results with those found previously, and limits of agreement (mean ± 1.96 SD). Bland–Altman plots were created to visually present the test–retest reliability. Test–retest reliability was analysed for three groups separately: all participants (n=184), participants who mentioned stable health during the past 4 weeks (n=135), and participants who mentioned health change during the past 4 weeks (n=39). A Mann–Whitney *U*-test was performed to investigate systematic differences on the mean score of

test and retest of the subscales, between participants who mentioned that their health changed during the past 4 weeks and participants who mentioned no health change during the past 4 weeks. The relation between the amount of health change mentioned and differences between pre-test and post-test was calculated using Spearman's correlations. Missing data were handled as prescribed by the manual of the RAND-36[50]. Statistical analyses were performed using SPSS 20.1 (IBM, New York, NY, USA). The level of significance was set at *p*-value of less than .05.

Results

In total, 276 (37%) patients completed the first questionnaire (table 1). Participants were statistically significantly younger compared with nonparticipants (t=2.815 (401.7); p=.005). The distribution of diagnoses statistically significantly differed between participants and nonparticipants (x^2=14.562 (7); p=.042). No statistically significant differences between participants and nonparticipants were found regarding sex (x^2=.718 (1); p=.397) and follow-up period (t=.667 (750); p=.505). Per scale 0 –4.7% of the items were missing. Scores on .4% (pain) to 3.6% (health change) of the subscales could not be calculated. Internal consistency (Cronbach's α) ranged from .81 (social functioning, vitality and general health) to .95 (physical functioning) (table 2). Floor and ceiling effects (>15%) were found for social functioning, role limitations – physical, role limitations – emotional, and pain (table 2).

Table 1: Personal characteristics of the participants, separated for the participants who completed only the first questionnaire (internal consistency) and both questionnaires (test-retest).

	Internal consistency (*n*=276)		Test-retest (*n*=184)	
	Mean	SD	Mean	SD
Age (years)	48.0	25.1	45.0	28.6
	Median	IQR	Median	IQR
Follow-up (months)	39	30 - 48	39.5	31 - 48
	n	%	*n*	%
Gender				
Men	148	54	106	58
Diagnosis group [a]				
Amputation	33	12	24	13
Cerebral Vascular Accident	46	17	31	17
Brain injury other [b]	35	13	23	12
Chronic pain	25	9	12	6
Multiple Sclerosis	32	12	25	14
Spinal cord injury	42	15	29	16
Other neurological disability [c]	34	12	22	12
Other [d]	29	11	18	10

Notes: SD = standard deviation, IQR = interquartile range ; [a] self-reported diagnosis; [b] brain injuries from traumatic or oncological origin and meningitis; [c] spina bifida, parkinson's disease and guillain-barré syndrome; [d] disabilities such as tumours, fibromyalgia, arthritis, multi trauma, chronic fatigue syndrome and decubitus ulcer.

Table 2: Internal consistency and dispersion, measured over the entire population (*n*=276)

	Internal consistency		Dispersion	
Subscale (number of items; score range)	α	Mean inter-item correlation	Floor (%)	Ceiling (%)
Physical Functioning (10; 20)	.95	.66	6.9	5.8
Social Functioning (2; 8)	.81	.68	2.5	22.1
Role Limitations – Physical (4; 4)	.90	.69	44.2	25.4
Role Limitations – Emotional (3; 3)	.88	.70	22.1	56.2
Mental Health (5; 25)	.85	.54	-	4.7
Vitality (4; 20)	.81	.52	-	1.1
Pain (2; 49)	.92	.86	1.1	27.2
General Health (5; 20)	.81	.47	.4	1.4
Health Change (1; 4)	-	-	7.2	3.6

Notes: α = Cronbach's alpha; internal consistency of the health change element cannot be determined since the scale consist of only one item.

Of 261 invited patients, 184 (70%) patients completed both questionnaires (table 1). The interval between the two tests ranged from 5 to 38 days (18.1 ± 5.3 days). The population that completed both questionnaires (test–retest population) was statistically significantly younger compared with the population that only completed the first questionnaire (t=3.551 (273.9); p<.001). The two populations did not differ statistically significantly on follow-up (t=−.716 (274); p=.475), diagnosis (x^2=6.823 (7); p=.448), and sex (x^2=3.526 (1); p=.060). No statistically significant differences were found between pre-test and post-test for the subscales (table 3). Test–retest reliability in ICC ranged from .71 (health change) to .88 (general health). Pearson r also ranged from .71 (health change) to .88 (general health) (table 3). Bland–Altman plots are presented in figure 1. Test–retest reliability is presented separately for the participants who mentioned health change during the past 4 weeks and the participants who indicated stable health (appendix 2). The participants who mentioned health change during the past 4 weeks scored statistically significantly lower on all subscale means of both questionnaires compared with the participants who mentioned a stable health during the past 4 weeks (U: 1387.50 to 1837.50, z: −4.408 to −2.659, p: <.001 to .008). The amount of health change was not statistically significantly related to the differences between pre-test and post-test on any of the subscales (Spearman's r ranged from .001 to .115; significance ranged from .140 to .991).

Table 3: Test-retest reliability of the RAND-36 (*n*=184)

Subscale (# of items; score range)	ICC (95% CI)	*r*	LoA	Test [a]	Retest [a]	Difference [a]	sig.
Physical Functioning (10; 20)	.86 (.82 ; .90)	.87	-36.4 ; 32.5	48.7 ± 31.7	50.2 ± 35.2	-1.9 ± 17.6	.357
Social Functioning (2; 8)	.75 (.68 ; .81)	.75	-39.2 ; 35.1	66.8 ± 27.2	68.9 ± 26.5	-2.0 ± 19.0	.125
Role Limitations – Physical (4; 4)	.79 (.73 ; .84)	.79	-57.4 ; 52.2	38.8 ± 42.0	42.1 ± 44.7	-2.6 ± 27.9	.301
Role Limitations – Emotional (3; 3)	.74 (.67 ; .80)	.74	-56.8 ; 62.6	70.0 ± 40.6	67.0 ± 44.1	2.9 ± 30.5	.251
Mental Health (5; 25)	.79 (.73 ; .84)	.79	-21.0 ; 22.7	74.1 ± 17.4	73.8 ± 17.6	.8 ± 11.2	.372
Vitality (4; 20)	.84 (.79 ; .88)	.84	-22.4 ; 24.0	57.0 ± 21.1	56.3 ± 21.1	.8 ± 11.8	.294
Pain (2; 49)	.87 (.83 ; .90)	.87	-29.8 ; 28.6	66.0 ± 29.0	66.2 ± 29.2	-.6 ± 14.9	.499
General Health (5; 20)	.88 (.84 ; .91)	.88	-23.3 ; 22.7	51.6 ± 23.5	51.8 ± 24.3	-.3 ± 11.7	.541
Health Change (1; 4)	.71 (.63 ; .77)	.71	-30.0 ; 32.2	45.9 ± 21.2	44.8 ± 20.4	1.1 ± 15.9	.220

Notes: # = number; ICC = Intraclass correlation coefficient; CI = Confidence interval; r = Pearson correlation; LoA = Limits of Agreement; sig = significance value of Wilcoxon signed ranks test; [a] mean ± SD

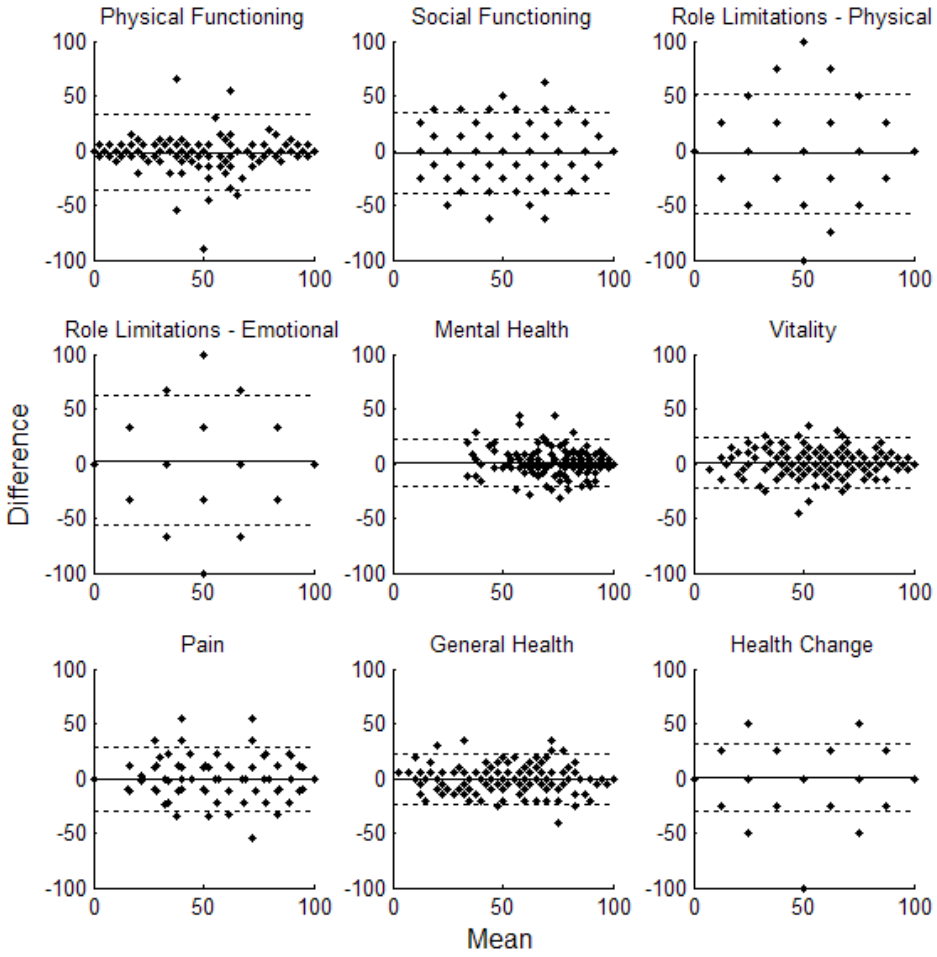

Figure 1: Bland-Altman plots for the individual subscales. Solid lines represent the mean difference between test and retest. Dotted lines represent the Limits of Agreement. (figure was created by using MATLAB 2014.b (The Mathworks Inc., Natrick, MA, USA))

Discussion

The aim of this study was to test the reliability of the RAND-36 in a post-rehabilitation population. The Dutch translation of the RAND-36 is reliable for research and group comparisons in a diverse post-rehabilitation population. However, it is not reliable for individual comparisons. Thirty-seven percent of the included patients completed the first questionnaire. This participation is lower compared with an earlier study in both community and chronic disease populations[67]. However, in that study the

investigators invited their population personally. Participation in our study was little higher compared with a study in healthy individuals, in which questionnaires were also sent by post[51]. Presumably, participation in the current study was related to the fact that participants were ex-patients, and familiar with our institute. Participation on the second questionnaire was relatively high (70%). Nonresponse could lead to nonresponse bias. Nonresponse bias is likely to influence RAND-36 outcomes, but the influence on reliability scores is not clear.

The criterion for internal consistency (Cronbach's α) is α more than or equal to .70[67]. All subscales fulfilled this criterion. Our findings were approximately equivalent to that of an earlier study on the Dutch translation of the RAND-36[51], except for the internal consistency of the social functioning scale, which was higher in our study. No remarkable differences between the internal consistency of subscales in our study and that in earlier research (SF-36 and RAND-36) were found, except for the pain scale, which was slightly higher in the current study[55–58,60–63].

In the current study, floor and ceiling effects were found on the subscales social functioning, role limitations – physical, role limitations – emotional, and pain. The floor and ceiling effects in both role limitations subscales may be explained by the small scoring range. These subscales consist of only a few dichotomous questions, resulting in a large proportion of patients with the same score. Floor and ceiling effects found in this study did not remarkably differ from that in earlier research[56,58,60–62].

No significant differences were found between pre-test and post-test for any of the subscales, indicating that HRQoL on all subscales remained constant between the two test occasions. In addition, for the participants who mentioned health change, and the participants who mentioned stable health during the past 4 weeks separately, no significant differences were found between pre-test and post-test (appendix 2). It seems remarkable that no differences between the test and retest were found within the participants who mentioned health change during the past 4 weeks. Moreover, the amount of health change mentioned is not related to differences between pre-test and post-test on any of the subscales. This means that although these participants experience health change, this is not detected by the subscales of the RAND-36. In six of nine subscales, the difference between test and retest is larger in the participants who mentioned health change compared with the participants who mentioned stable health. The absence of a significant difference between test and retest is suggested to be explained by the low number of participants who mentioned health change.

For group comparisons and research, a minimum ICC of .70 is required[68]. All subscales

fulfilled this criterion. For individual comparisons, a minimum ICC of .90 is required[68]. None of the subscales fulfilled this criterion, whereby none of the subscales is reliable enough for individual comparisons in this population. Compared with earlier research regarding different translations of both RAND-36 and SF-36[52,56–63], the current study shows similar test–retest reliability (ICC) on all subscales (appendix 1). Compared with a study that assessed reliability of the Dutch translation of the RAND-36 in a brain injury population[52], reliability was similar or higher in our study for all subscales, except for role limitations – emotional, which scored much lower in our study. However, the ICC of the role limitations – emotional scale in the study of Van Baalen et al. (2006) is relatively high compared with the results of other studies, and compared with other subscales in their study. It can be hypothesised that this result might be an outlier. The higher reliability on most subscales in our study may be explained by the population tested. The current study tested a more heterogeneous post-rehabilitation population, whereas the earlier study tested a relatively homogeneous brain injury population[52]. The ICC is calculated as the ratio of the between-subjects variation and the total (between + within participants) variation. In a very heterogeneous population, between-subjects variation will be high; hence, the influence of the within-participants variation will be smaller.

Although Pearson *r* cannot detect systematic error, it was calculated in order to compare our findings with the manual belonging to the Dutch translation of the RAND-36[50]. When comparing with earlier research in healthy adults[51], the current study found higher test–retest reliability (Pearson *r*) for all subscales (appendix 1). The higher test–retest reliability found in the current study may be explained by the shorter interval between the test and retest compared with the study of Van der Zee et al. (1993) (appendix 1). When comparing our findings with findings in a diverse population including low back pain, reliability of all subscales was comparable[55].

In figure 1, Bland–Altman plots are provided to visually present the agreement of the subscales. Figure 1 shows that the subscales with the highest ICCs (general health (.88), pain (.87) and physical functioning (.86)) have small limits of agreement. Within both role limitations subscales, broad limits of agreement are found, whereas ICC values are acceptable. The relatively high ICCs are caused by a high between-subjects variation, owing to a heterogeneous population. SDs of both test and retest scores are remarkably high for the role limitations – physical and role limitations – emotional subscales (table 3).

When testing the reliability of the subscales for participants who mentioned health change during the past 4 weeks and participants who mentioned stable health

separately, it is remarkable that the reliability of both role limitations subscales is lower in participants with stable health compared with participants who mentioned health change (appendix 2). The low reliability of both role limitations subscales can be explained by the small scoring range (difference between minimal and maximal possible raw score). In scales that have a small scoring range, different answers result in highly different scale scores, which impairs test–retest reliability[68]. However, no explanation could be found for the higher reliability in the health change group compared with the stable health group. Notably, participants who mentioned health change during the past 4 weeks scored lower on all subscales (mean score of test and retest) compared with participants who mentioned stable health (appendix 2). This suggests that stability of health leads to a higher quality of life.

A strength of the current study is the sample size, which is relatively large compared with earlier reliability studies (appendix 1). This study is limited by the variation on the interval between the two test occasions between participants. An additional check was done to ensure stability of the construct under measurement. However, the shortest interval between test and retest was 5 days, whereby it could have been that the participant remembered the given answers.

Clinically, the results of this study indicate that the RAND-36 is a reliable instrument for performing research and group comparisons in rehabilitation. The RAND-36 is suitable for describing the HRQoL of diverse patient groups that were represented in the post-rehabilitation population tested in the current study. Moreover, it can be used to assess differences between HRQoL of different populations, and for evaluation of healthcare or interventions, provided that it will be evaluated in a whole group. However, the RAND-36 is not appropriate for individual HRQoL assessment, or monitoring individual progression in rehabilitation. This finding agrees with the suggestion given by an earlier review on different health status surveys[65]. However, the current study more elaborates test–retest reliability by using ICC instead of Pearson *r*, and illustrates the findings by using Bland–Altman plots.

Conclusions

The Dutch translation of the RAND-36 is a reliable instrument for measuring HRQoL in a diverse post-rehabilitation population, when used for group comparisons or research. Overall, the Dutch translation of the RAND-36 was proven to be not reliable enough for individual comparisons in a diverse post-rehabilitation population.

Appendix 1: Test-retest reliability of the RAND-36 and SF-36, findings of the current study and earlier research

	Test	Population	n	Int.	PF	SF	RP	RE	MH	VT	BP	GH	HC
ICC													
Current study	*RAND (d)*	*Rehabilitation*	*184*	*2*	*.86*	*.75*	*.79*	*.74*	*.79*	*.84*	*.87*	*.88*	*.71*
Van Baalen et al. (2007)	RAND (d)	Brain injury	14	2	.64	.80	.66	.94	.44	.81	.67	.68	-
Dorman et al. (1998)	SF (e)	CVA	209	3	.80	.79	.77	.60	.30	.77	.81	.81	-
Lin et al. (2007)	SF (c)	SCI	10	2	.71	.93	.89	.99	.77	.93	.87	.85	-
Steffen et al. (2008)	SF (e)	Parkinsonism	36	1	.80	.71	.85	.84	.83	.88	.89	.85	-
Ferrer et al. (2006)	SF (s)	CLBP	43	2-13	.91	.77	.70	.55	.71	.75	.81	.82	-
Boyer et al. (2006)	SF (f)	HNMD	69	2	.75	.63	.73	.71	.78	.85	.82	.78	-
Dunn et al. (2003)	SF (e)	Back pain	14	2	.93	.88	.81	.74	.90	.94	.89	.96	-
Saleh et al. (2002)	SF (e)	Hip/knee pain	36	3	.90	.85	.91	.79	.93	.86	.87	.84	-
Vickrey et al. (1997)	SF (e)	MS	84	2	.96	.64	.66	.72	.85	.82	.86	.71	-
Pearson *r*													
Current study	*RAND (d)*	*Rehabilitation*	*184*	*2*	*.87*	*.75*	*.79*	*.74*	*.79*	*.84*	*.87*	*.88*	*.71*
Van der Zee et al. (1993)	RAND (d)	Healthy	159	9	.82	.58	.60	.67	.73	.76	.72	.80	.40
Ruta et al. (1994)	SF (e)	Diverse*	414	2	.93	.80	.76	.66	.81	.84	.82	.88	-

Notes: Int. = Mean Interval between test and retest (weeks); PF = physical functioning; SF = social functioning; RP = role limitations – physical; RE = role limitations – emotional; MH = mental health; VT = vitality; BP = bodily pain; GH = general health; HC = health change; ICC = Intraclass Correlation Coefficient; r = Pearson correlation; CVA = Cerebral Vascular Accident; SCI = Spinal Cord Injury; CLBP = Chronic Low Back Pain; HNMD = Hereditary Neuromuscular Disease; MS = Multiple Sclerosis; d = Dutch translation, e = English translation, c = Chinese translation, s = Spanish translation; f = French translation; * Diverse population including low back pain, menorrhagia

Appendix 2: Test-retest reliability of the Dutch RAND-36 separated for participants that indicated health change, and participants that indicated stable health during the past four weeks

Table A2.1: Test-retest reliability of the RAND-36 for participants who mentioned stable health during the past 4 weeks (*n*=135)

Subscale (# of items; score range)	ICC (95% CI)	*r*	LoA	Test [a]	Retest [a]	Difference	Sig.
Physical Functioning (10; 20)	.87 (.82 ; .90)	.87	-36.1 ; 31.9	52.9 ± 31.3	54.4 ± 35.5	-2.1 ± 17.3	.270
Social Functioning (2; 8)	.80 (.73 ; .85)	.80	-34.7 ; 29.7	71.1 ± 26.9	73.6 ± 25.6	-2.5 ± 16.4	.060
Role Limitations – Physical (4; 4)	.79 (.72 ; .85)	.79	-56.7 ; 53.6	45.7 ± 42.5	48.1 ± 44.7	-1.6 ± 28.1	.568
Role Limitations –Emotional (3; 3)	.67 (.56 ; .76)	.68	-56.4 ; 63.7	78.2 ± 34.7	74.3 ± 40.4	3.7 ± 30.6	.210
Mental Health (5; 25)	.80 (.72 ; .85)	.80	-19.5 ; 20.8	77.2 ± 15.8	76.4 ± 16.5	.6 ± 10.3	.463
Vitality (4; 20)	.83 (.76 ; .87)	.83	-23.2 ; 23.5	60.4 ± 20.6	60.3 ± 19.7	.1 ± 11.9	.740
Pain (2; 49)	.87 (.81 ; .90)	.87	-29.2 ; 26.7	71.7 ± 28.2	72.5 ± 26.7	-1.2 ± 14.3	.321
General Health (5; 20)	.88 (.84 ; .91)	.88	-22.4 ; 22.0	55.3 ± 22.9	55.4 ± 23.9	-.2 ± 11.3	.845
Health Change (1; 4)	.73 (.64 ; .80)	.73	-24.5 ; 25.6	48.7 ± 17.9	48.1 ± 16.9	.6 ± 12.8	.607

Notes: # = number; ICC = Intraclass correlation coefficient; CI = Confidence interval; *r* = Pearson correlation; LoA = Limits of Agreement; sig = significance value of Wilcoxon signed ranks test; [a] mean ± SD

Appendix 2 (continuation)

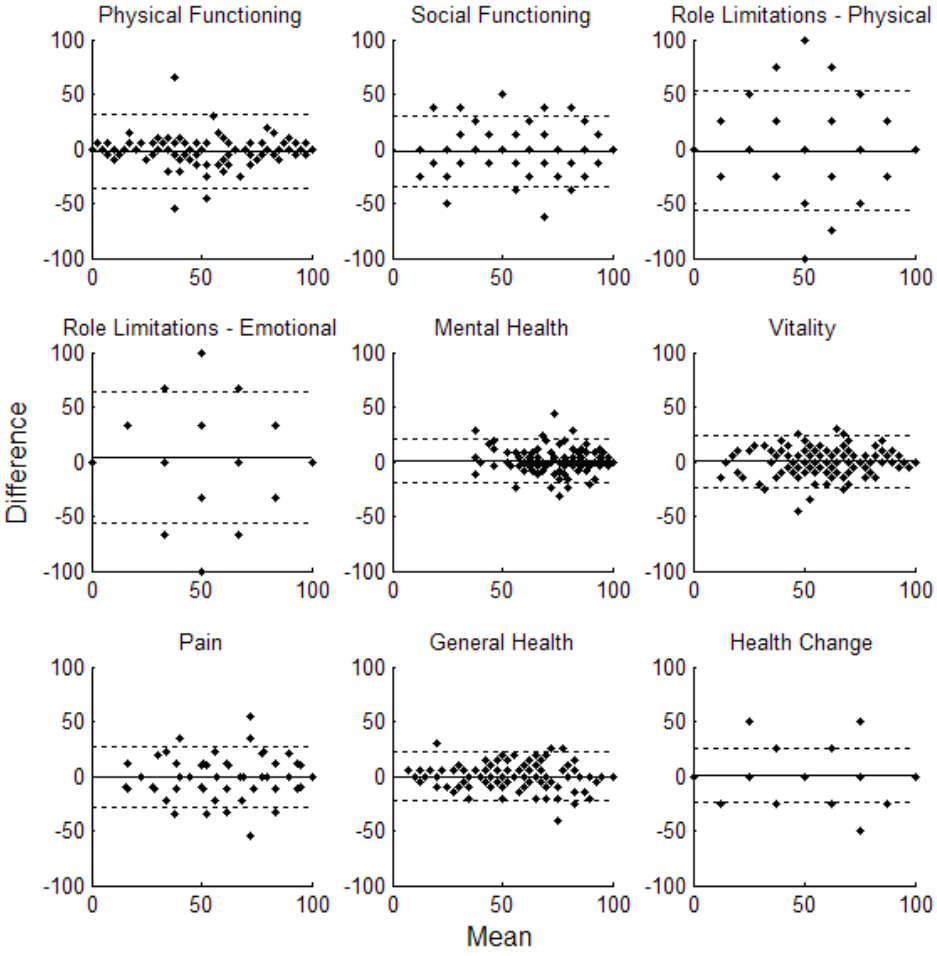

Figure A2.1: Bland-Altman plots for the individual subscales of participants who mentioned stable health during the past 4 weeks (*n*=135). Solid lines represent the mean difference between test and retest. Dotted lines represent the Limits of Agreement. (figure was created by using MATLAB 2014.b (The Mathworks Inc., Natrick, MA, USA))

Appendix 2 (continuation)

Table A2.2: Test-retest reliability of the RAND-36 for participants who mentioned health change during the past 4 weeks (*n*=39)

Subscale (# of items; score range)	ICC (95% CI)	*r*	LoA	Test [a]	Retest [a]	Difference	Sig.
Physical Functioning (10; 20)	.86 (.75 ; .92)	.86	-35.7 ; 30.3	36.8 ± 30.6	39.5 ± 32.8	-2.7 ± 16.9	.837
Social Functioning (2; 8)	.36 (.05 ; .61)	.36	-52.7 ; 51.5	55.8 ± 24.8	56.4 ± 22.0	-.6 ± 26.6	.887
Role Limitations – Physical (4; 4)	.92 (.85 ; .96)	.92	-29.7 ; 28.4	20.5 ± 35.3	21.7 ± 37.3	-.7 ± 14.8	.792
Role Limitations –Emotional (3; 3)	.79 (.62 ; .88)	.78	-58.8 ; 64.1	48.7 ± 46.4	45.6 ± 48.0	2.6 ± 31.4	.571
Mental Health (5; 25)	.72 (.52 ; .84)	.72	-25.6 ; 29.6	68.0 ± 18.1	66.1 ± 19.0	2.0 ± 14.1	.528
Vitality (4; 20)	.82 (.68 ; .90)	.83	-20.5 ; 26.5	48.4 ± 19.4	45.1 ± 20.7	3.0 ± 12.0	.176
Pain (2; 49)	.80 (.65 ; .89)	.80	-33.0 ; 35.5	50.7 ± 25.1	49.5 ± 29.3	1.3 ± 17.5	.799
General Health (5; 20)	.83 (.69 ; .91)	.83	-27.4 ; 25.8	40.0 ± 23.3	39.7 ± 23.1	-.8 ± 13.6	.420
Health Change (1; 4)	.64 (.41 ; .80)	.64	-45.0 ; 47.6	37.2 ± 28.0	35.9 ± 27.4	1.3 ± 23.6	.416

Notes: # = number; ICC = Intraclass correlation coefficient; CI = Confidence interval; *r* = Pearson correlation; LoA = Limits of Agreement; sig = significance value of Wilcoxon signed ranks test; [a] mean ± SD

Appendix 2 (continuation)

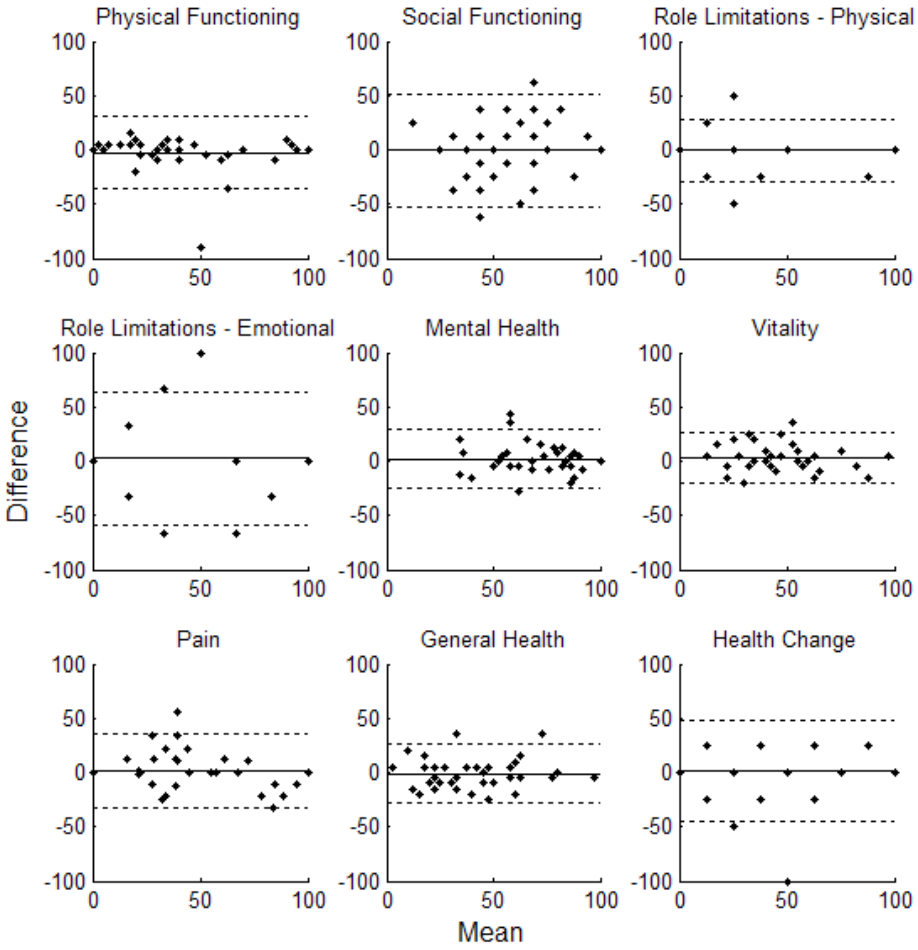

Figure A2.2: Bland-Altman plots for the individual subscales of participants who mentioned health change during the past 4 weeks (*n*=39). Solid lines represent the mean difference between test and retest. Dotted lines represent the Limits of Agreement. (figure was created by using MATLAB 2014.b (The Mathworks Inc., Natrick, MA, USA))