INTRA- AND INTER-RATER RELIABILITY 'FUNCTIONAL INFORMATION SYSTEM' AND 'FUNCTIONAL ABILITY LIST'

CHAPTER 5

Sandra Brouwer
Pieter U. Dijkstra
Erwin H.J. Gerrits
Jan M.H. Schellekens
Johan W. Groothoff
Jan H.B. Geertzen
Ludwig N.H. Göeken

This chapter is a translation of : Intra- en inter-beoordelaarsbetrouwbaarheid 'FIS-belastbaarheidspatroon' en 'Functionele Mogelijkheden Lijst' Tijdschrift voor Bedrijfs- en Verzekeringsgeneeskunde 2002;11:360-367.
Reprinted with the kind permission of TBV.

ABSTRACT

Objective: To investigate the reliability of the Functional Information System (FIS) and the Functional Ability List (FAL) in the assessment of work limitations in chronic low back pain (CLBP) patients in rehabilitation treatment.
Design: To assess the intra-rater reliability a social insurance physician filled out the FIS and FAL twice after history taking and physical examination of CLBP patients with a two-week interval. To assess the inter-rater reliability, two social insurance physicians performed the diagnostic procedure independently and filled out the FIS and FAL. The first physician carried out the history taking and performed the physical examination, and the next day the other physician did the same. Each subject was examined 4 times. A Kappa value of more than 0.60 was considered to be acceptable.[11] The predetermined interpretation of the percent age of agreement was arbitrarily set on 80%.
Subjects: Thirty patients with chronic low back pain, referred for treatment.
Results: Unacceptable intra- and inter-observer reliability for almost all items of the FIS were found, kappa-values and percentage agreement were below the criteria for acceptance. The FAL showed better results, higher kappa-values as well as higher percentage agreement. However, for a great part of the items, kappa-values and percentage agreement were still unacceptable.
Conclusion: FIS and FAL are not reliable instruments for assessing work limitations in CLBP-patients, and therefore not useful as instruments in Rehabilitation Medicine.

INTRODUCTION

In Rehabilitation Medicine, aim of treatment in patients with chronic low back pain (CLBP) is to optimise daily functioning by limitation reduction. Several questionnaires are in use to assess treatment effectiveness, of which the Roland Morris Disability Questionnaire (RMDQ) is one of the most frequently used questionnaires. The RMDQ proved to be reliable and valid[1-6] to assess perceived limitations in 24 activities of daily living (ADL). CLBP however, has a considerable impact on patient's performance of activities in the work-situation. To assess the severity of work limitations, an instrument developed to assess these limitations should be used. However, in Rehabilitation Medicine, these work-related instruments are absent.

In Dutch Social Insurance Medicine, the work limitations are assessed on the basis of a history taking, a physical examination and relevant information from other disciplines (for example from company physicians). To standardise the procedure and to quantify the findings, the Functional Information System (FIS) was used as scoring form by social insurance physicians for many years. The FIS consists of 15 physical items, 12 vocational items and 1 psychological item. Most items have an ordinal rating scale with 8 to10 ratings increasing in duration or frequency of the activity.

Since 2000, the FIS has been replaced by a new scale: the ‘Function Ability List (FAL), developed by the 'Claim Beoordelings- en Borgingssysteem' (CBBS). The FAL consists of 6 modules: 1. personal functioning, 2. social functioning, 3. adaptation to environmental demands 4. dynamic performance 5. static posture 6. working hours. The modules 3 to 6 include more or less of the same items as the FIS. Some of the FIS items are broken down into sub items (for example reaching is divided into 1. reaching (maximal distance in cm), 2. long-time and frequently reaching (150-500 times an hour). In general, the FAL uses less specific rating scales (2-3 points scale) than the FIS.
The social insurance physicians assess patients' work limitations by means of history taking and physical examination. The FIS and FAL are standardised assessment methods, used to describe the findings in the same manner and used as a communication instrument between the social insurance physician and the occupational specialist.

In Rehabilitation Medicine, there is a need for instruments to assess the work limitations of CLBP patients. To know whether the FIS and FAL can be used as work-related instruments, the qualities of these instruments should be investigated in rehabilitation practice. Demonstration of the reliability is a minimum requirement. Reliability is a comprehensive term and can be investigated in several ways.

In this study, the reliability of observers is relevant, intra- and inter-observers reliability is investigated. The intra-rater reliability refers to which extent one observer, who assesses the same subjects repeatedly under the same circumstances, obtains similar results. The inter-rater reliability refers to which extent two or more observers, who assess the same subjects under the same circumstances, obtain similar results.[7]
The intra- and inter-observer reliability of the FIS proved to be moderate to good.[8] Critical analysis of that study, however, showed that the used procedure differs from the procedure in daily practice. Video-tapes of the history taking and physical examination of 14 patients performed by one physician were used. Twenty-two physicians assess the work limitations of the patients on the basis of those video-tapes and filled out the FIS. Using video-tapes excludes interactions between patients and physicians who fill out the FIS. It is debatable whether those study results can be generalised to daily practice. The reliability of the FAL has not been investigated yet.

Aim of the study is to investigate the intra- and inter observer reliability of the FIS and FAL in the assessment of work limitations. The procedure in the present study corresponds to the procedure in daily practice.

METHOD

Subjects
Thirty patients with CLBP (24 men, 6 women) who were referred for treatment to a rehabilitation centre from May 2000 to April 2001. The mean age of the patients was 40 years (SD 8.1). The median duration of low back pain ranged between 5 and 10 years. None of the patients had had surgery for their back pain. Patients were only included, if they were still at work or less than one year out of work. Patients were out of work with a mean of 17 weeks (SD 19.2). Half of the patients were receiving Workers' compensation.

Physicians
Four registered Dutch social insurance physicians (3 men, 1 woman) enrolled the study. The average length of time spent in professional practice was 15 years. Three of the physicians are employed at 'Uitvoering Werknemers-Verzekeringen' (UWV) and 1 physician is employed at Argonaut.

Procedure
Patients were recruited from the population who was admitted for rehabilitation treatment of the Centre for Rehabilitation at the University Hospital Groningen. They were given a description of the study and all patients were assessed during a visit to the occupational assessment centre of the Centre of Rehabilitation Beatrixoord in Haren, the Netherlands. Demographic information and low back

pain history were obtained of all patients. Prior to this study, the social insurance physicians had achieved consensus on the assessment of the items. Furthermore, both history taking and physical examination were standardised. Thirteen of the 15 items of the FIS were assessed. Two items were excluded, because it was felt that these items could not be limited due to low back pain: neck use, and feeling and fingering. In this study, the first version of the FAL was used.
The study consisted of two sessions. The first session took place within two days (testing day 1 and 2). Two weeks separated the first en second session. The second session took also place within two days also (testing day 3 and 4). Each patient was assessed by two social insurance physicians. The first physician carried out the history taking and performed the physical examination, and the next day the other physician did the same. After the assessment, each physician independently determined patient's work limitations and filled out the FIS and FAL. After two weeks, the same physicians in the same order performed the assessment again and filled out the FIS and FAL. Thus, each subject was examined 4 times enabling analyses of intra- and inter-observer reliability. Time and day and place of assessment were held constant. All patients were assessed prior to entering a rehabilitation program.

Analysis
Two different types of measures of reliability were used to determine the reliability of the FIS and FAL. First, Kappa values were calculated. In case of dichotomous outcome Cohen's Kappa[9] was calculated and in case of ordinal outcome a weighted kappa was calculated.[10] Second, the percentage of absolute agreement was calculated. A Kappa value of more than 0.60 was considered to be acceptable.[11] The predetermined interpretation of the percent absolute agreement was arbitrarily set on 80%.
To visualise the variation of the results the percentage of agreement of scores with a difference of 1 point between both sessions was calculated also. Furthermore, the percentage of agreement of scores with a difference of ≥ 2 points was calculated.
The statistical software package Agree 7.0[12] was used for the calculation of Kappa values and the statistical software package SPSS (version 10) was used for percentage agreement calculations.

RESULTS

Intra-observer reliability of FIS and FAL of 49 cases was measured. Eleven cases were not available, because patients did not feel capable to attend the second session (2 cases), or the same combination of physicians could not be made for the second session (9 cases).
Inter-observer reliability of FIS and FAL of 29 cases was measured in the first session. One case was not available due to absence of the second physician. In the second session, inter-observer reliability of 24 cases was measured. Six cases were not available, due to the absence of the second physician.

Intra-observer reliability FIS
Kappa values of the intra-observer reliability ranged from 0.43 to 0.74 (table 1). Only 3 of the 13 items (23%) reached the acceptable level of 0.60. The percentages of absolute agreement ranged from 43 to 86 percentages. Only 1 item reached the level of 80%.

Inter-observer reliability FIS
Kappa values of the inter-observer reliability study on session 1 ranged from -0.16 to 0.24. The percentages of absolute agreement ranged from 23 to 57% (table 2). In session 2, kappa values from –0.13 to 0.35, and the percentage of absolute agreement ranged from 13 to 56% (table 3). None of the values reached the acceptable level of 0.60 (kappa) or 80% (percentage agreement).

Variation FIS
The variation of results within observers (intra-observer reliability) of the FIS is shown in table 1, column 4-5. In 14-46% the results differed 1 point, which means that the results of one observer on both sessions differ 1 point. In 0-28%, the results differed ≥ 2 points. Between observers (inter-observer reliability) in session 1, 27-60% of the results showed a difference of 1 point, in 0-43% the results differed ≥ 2 points. (table 2, column 4-5). In session 2, 30-50% of the results showed a difference of 1 point and 0-50% a difference of ≥ 2 points (table 3, column 4-5).

Table 1. Intra-observer reliability of FIS (49 paired observations)

Activities	Kappa	Absolute agreement (%)	Difference of 1 (%)	Difference of ≥ 2 (%)	Reliability
Working above shoulder level	0.43	49	46	5	N
Sustained and frequent bending & rotating	0.52	59	25	16	N
Carrying	0.74	76	18	6	A
Push and pull static*	0.66	86	14	0	A
Working static forward bending	0.50	53	35	12	N
Climbing and clambering	0.51	45	45	10	N
Kneeling, crawling, squatting	0.45	53	37	10	N
Walking	0.58	57	35	8	N
Reaching	0.43	55	41	4	N
Standing	0.55	45	27	28	N
Lifting	0.72	67	31	2	A
Stair climbing	0.54	57	39	4	N
Sitting	0.53	71	27	2	N

* Dichotomous data (Cohen's Kappa). Other items are ordinal data (weighted Kappa)

A: Acceptable

N: Not acceptable

Table 2. Inter-observer reliability FIS, session 1 (n=30)

Activities	Kappa	Absolute agreement (%)	Difference of 1 (%)	Difference of ≥ 2 (%)	Reliability
Working above shoulder level	0.24	43	47	10	N
Sustained and frequent bending & rotating	0.13	37	27	36	N
Carrying	0.22	33	43	24	N
Push and pull static*	-0.16	50	50	0	N
Working static forward bending	-0.01	23	50	27	N
Climbing and clambering	0.12	30	30	40	N
Kneeling, crawling, squatting	0.15	33	40	27	N
Walking	0.23	43	30	27	N
Reaching	0.00	40	43	17	N
Standing	0.16	30	27	43	N
Lifting	0.24	27	53	20	N
Stair climbing	0.18	30	60	10	N
Sitting	0.23	57	40	3	N

* Dichotomous data (Cohen's Kappa). Other items are ordinal data (weighted Kappa)

N: Not acceptable

Table 3. Inter-observer reliability of FIS, session 2 (n=24)

Activities	Kappa	Absolute agreement (%)	Difference of 1 (%)	Difference of ≥ 2 (%)	Reliability
Working above shoulder level	0.19	38	50	12	N
Sustained and frequent bending & rotating	0.20	39	30	21	N
Carrying	0.29	38	33	29	N
Push and pull static*	-0.12	50	50	0	N
Working static forward bending	-0.13	12	38	50	N
Climbing and clambering	0.23	38	42	20	N
Kneeling, crawling, squatting	0.28	38	50	12	N
Walking	0.48	46	42	12	N
Reaching	0.15	42	42	16	N
Standing	0.35	21	50	29	N
Lifting	0.23	29	38	33	N
Stair climbing	0.10	33	46	21	N
Sitting	0.40	56	44	0	N

* Dichotomous data (Cohen's Kappa). Other items are ordinal data (weighted Kappa)

N: Not acceptable

Intra-observer reliability FAL
Kappa values of the intra-observer reliability ranged from 0.00 to 0.75 (table 4). Ten of the 26 items (39%) reached the acceptable level of 0.60. The percentage of absolute agreement of the FAL ranged from 47 to 96%. Fifteen of the 26 items (58%) reached the level of 80%.

Inter-observer reliability FAL
Kappa values of the inter-observer reliability study in session 1 ranged from –0.46 to 0.75. Two of the 26 items (8%) reached the acceptable level of 0.60. The percentage of absolute agreement ranged from 31 to 97% (table 5). Three of the 26 items (12%) reached the level of 80%.
In session 2, kappa values ranged from –0.30 to 1.00 (table 6). Two of the 26 items (8%) reached the acceptable level of 0.60. The percentage of absolute agreement ranged from 38 to 100%. Four of the 26 items (16%) reached the level of 80%.

Variation FAL
The variation of results within observers (intra-observer reliability) of the FAL is shown in table 5, column 4-5. In 4-53%, the results differed 1 point, in 0-3% the results differed ≥ 2 points. Between observers (inter-observer reliability) in session 1, 0-59% of the results showed a difference of 1 point, in 0-17% the results differed ≥ 2 points (table 5, column 4-5). In session 2, 0-62% of the results showed a difference of 1 point and 0-9% showed a difference of ≥ 2 points (table 6, column 4-5).

DISCUSSION AND CONCLUSION

The results showed unacceptable intra- and inter-observer reliability for almost all items of the FIS based upon kappa values (0.60 or higher) or based upon percentage absolute agreement (80% or higher). The FML showed better results, higher kappa values as well as higher percentage absolute agreement. However, for a great part of the items, kappa values and percentage absolute agreement were below the criteria of acceptance.

These insufficient reliability results may be caused by the variation of the patient and of physicians. Patients may not report their problems in the same way to both physicians or in both sessions, or may emphasise different sub problems during the different examinations. Factors like motivation, tiredness, pain and memory may influence the variation at the moment of examination. Another source of variation may be the physician. Each physician has his own standardised way of examination. Furthermore, motivation, tiredness and memory of the physicians may play a role in causing variation. Other reasons for variation may be a possible learning effect of repeated assessment for both the patients as well as the physicians. Because these sources of variation may play a role in daily practice also, we decided not to exclude these sources by using for example video-tapes.

A limitation of the study was that not all combinations of physicians could be made in the same frequency. Ideally, the frequency of combinations of each physician to the other three physicians should be equal, furthermore the frequency of being first or second observer should be equal. In this study, this ideal situation could not be reached. The availability of the physicians determined the combinations. Due to two fixed testing days of the week and the availability of some of the physicians only on one specific day, some of the combinations could not be made and some of the combinations were made more often.

The reliability of the FAL proved to be higher than the reliability of the FIS. This difference may be caused by differences in number of response categories each item. The FAL consists of 2 to 3 categories each item, the FIS consists of 8 to10 categories each item. The more response categories the higher the chance of variation in responses within and between observers and within patients. The variation of the responses within and between observers on the FIS and FAL confirmed this assumption: the number of items with absolute agreement or with a variation of 1 on the FAL was larger than the number of items with absolute agreement or with a variation of 1 on the FIS.
Another possible cause of higher reliability on the FAL than on the FIS may be stricter definitions of the response categories of the FAL, which may restrict the interpretation margin of the physicians.

Table 4. Intra-observer reliability of FAL (45 paired observations)

Activities	Kappa (weighted)	Absolute agreement (%)	Difference of 1 (%)	Difference of ≥ 2 (%)	Reliability
A) Body movements scale					
Reaching (cm)	0.00	96	4	0	A
Sustained and frequent reaching (150-500 times each hour)	0.58	87	13	0	N
Bending (degrees)	0.71	76	24	0	A
Sustained and frequent bending (150-500 times each hour)	0.59	76	24	0	N
Rotation*	0.71	87	13	0	A
Push and pull static	0.64	84	16	0	A
Lifting or carrying (1-15 kg)	0.44	64	33	3	N
Sustained and frequent lifting or carrying light (150-500 times per hour)	0.44	71	29	0	N
Sustained and frequent lifting or carrying heavy (50 times each hour)	0.65	80	20	0	A
Walking (0-30 min)	0.45	82	16	2	N
Sustained walking (part of a day)*	0.65	82	18	0	A
Stair climbing	0.41	69	29	2	N
Climbing	0.61	82	18	0	A
Kneeling or squatting*	0.43	87	13	0	N
Other limitations in body movements*	0.60	89	11	0	A
Special work requirement*	0.75	88	12	0	A

B) Body postures scale					
Sitting (0-60 min)	0.51	73	37	0	N
Sitting prolonged (part of a day)*	0.51	76	24	0	N
Standing (0-30 min)	0.42	47	53	0	N
Standing prolonged (part of a day)*	0.69	93	7	0	A
Kneeling or squatting*	0.32	71	29	0	N
Bend or rotated*	0.46	73	27	0	N
Working above shoulder level*	0.20	76	24	0	N
Alternating postures*	0.64	82	18	0	A
Other limitations in body movements*	0.29	91	9	0	A
Special work requirements*	0.54	80	20	0	N

* Dichotomous data (Cohen's Kappa). Other items are ordinal data (weighted Kappa)

A: Acceptable

N: Not acceptable

Table 5. Inter-observer reliability of FAL, session 1 (n=29)

Activities	Kappa	Absolute agreement (%)	Difference of 1 (%)	Difference of ≥ 2 (%)	Reliability
A) Body movements scale					
Reaching (cm)	-0.06	83	17	0	A
Sustained and frequent reaching (150-500 times each hour)	-0.46	66	31	3	N
Bending (degrees)	0.75	79	21	0	A
Sustained and frequent bending (150-500 times each hour)	0.00	52	41	7	N
Rotation*	-0.02	55	45	0	A
Push and pull static	-0.10	52	45	3	A
Lifting or carrying (1-15 kg)	0.44	62	38	0	N
Sustained and frequent lifting or carrying light (150-500 times per hour)	0.34	72	28	0	N
Sustained and frequent lifting or carrying heavy (50 times each hour)	0.35	72	21	7	A
Walking (0-30 min)	0.03	66	34	0	N
Sustained walking (part of a day)*	0.29	62	38	0	A
Stair climbing	0.16	59	38	3	N
Climbing	0.24	59	41	0	A
Kneeling or squatting*	-0.12	79	21	0	N
Other limitations in body movements*	0.20	79	21	0	A
Special work requirement*	0.16	59	41	0	A

B) Body postures scale					
Sitting (0 -60 min)	0.30	66	34	0	N
Sitting prolonged (part of a day)*	0.08	59	38	3	N
Standing (0-30 min)	0.01	31	52	17	N
Standing prolonged (part of a day)*	0.35	90	10	0	A
Kneeling or squatting*	-0.25	41	59	0	N
Bend or rotated*	0.23	62	38	0	N
Working above shoulder level*	-0.25	59	41	0	N
Alternating postures*	0.19	52	48	0	A
Other limitations in body movements*	0.65	97	3	0	A
Special work requirements*	0.23	62	38	0	N

* Dichotomous data (Cohen's Kappa). Other items are ordinal data (weighted Kappa)

A: Acceptable

N: Not acceptable

Table 6. Inter-observer reliability of FAL, session 2 (n=24)

Activities	Kappa	Absolute agreement (%)	Difference of 1 (%)	Difference of ≥ 2 (%)	Reliability
A) Body movements scale					
Reaching (cm)	1.00	100	0	0	A
Sustained and frequent reaching (150-500 times each hour)	-0.06	63	37	0	N
Bending (degrees)	0.76	83	17	0	A
Sustained and frequent bending (150-500 times each hour)	0.52	71	29	0	N
Rotation*	0.29	67	33	0	N
Push and pull static	-0.03	58	42	0	N
Lifting or carrying (1-15 kg)	0.38	63	37	0	N
Sustained and frequent lifting or carrying light (150-500 times per hour)	0.32	67	33	0	N
Sustained and frequent lifting or carrying heavy (50 times each hour)	0.45	67	33	0	N
Walking (0-30 min)	0.19	75	21	4	N
Sustained walking (part of a day)*	0.58	79	21	0	N
Stair climbing	0.14	58	33	9	N
Climbing	0.28	71	29	0	N
Kneeling or squatting*	-0.11	79	21	0	N
Other limitations in body movements*	-0.08	67	33	0	N
Special work requirement*	-0.18	38	62	0	N

B) Body postures scale					
Sitting (0 -60 min)	0.29	63	33	4	N
Sitting prolonged (part of a day)*	0.53	75	25	0	N
Standing (0-30 min)	0.34	54	38	8	N
Standing prolonged (part of a day)*	0.47	92	8	0	N
Kneeling or squatting*	0.14	67	33	0	N
Bend or rotated*	0.18	58	42	0	N
Working above shoulder level*	0.11	67	33	0	N
Alternating postures*	-0.17	42	58	0	N
Other limitations in body movements*	-0.04	92	8	0	A
Special work requirements*	-0.30	46	54	0	N

* Dichotomous data (Cohen's Kappa). Other items are ordinal data (weighted Kappa)

A: Acceptable

N: Not acceptable

To calculate reliability, different indices of reliability were used. Cohen's Kappa was used for the dichotomous data and weighted kappa for ordinal data, percentage absolute agreement was used for all data.
The usefulness of Kappa as the only measure of reliability is problematic in case of lack of variation in cell fillings. In our study there is a large proportion of agreement, most of which is limited to only one of the possible rating choices. Under this limited variation, excellent absolute agreement can be found whereas the Kappa values do not reach the level of acceptance (for example, reaching and other limitations in body postures in table 4). Furthermore, only one decision can make the difference between poor and excellent reliability expressed as a Kappa. For example, the inter-observer reliability expressed as a Kappa in session 1 of reaching was -0.06, in session 2 the Kappa value was 1.00. In these cases, it is better to use percentage of absolute agreement. An important weakness of this simple calculation is that it does not take into account the agreement that is expected to occur due to chance alone.[9,13] Kappa corrects the observed agreement for agreement that is expected by chance.
Because both indices of statistical analyses have their strengths and limitations, it was decided to apply both statistical analyses in this study in contrast to the study of Spanjer[8], which reported the variation and absolute agreement only.

In conclusion, based on the study results, the FIS and FAL are not reliable instruments for the assessment of work limitations in CLBP patients. It is debatable whether these instruments are useful in daily practice. In Rehabilitation Medicine, the choice of the usefulness of an instrument depends more and more on the psychometric properties of the instrument (reliability, validity and responsiveness). If the FIS and FAL could be used in Rehabilitation Medicine, the psychometric properties should be sufficient. Based on the reliability results it can be concluded that both FIS and FAL are not useful instruments in Rehabilitation Medicine.
Whether the results of this research can be applied to the Social Insurance Medicine is unclear. This study took place within Rehabilitation Medicine, in CLBP patients who volunteered to participate were less than 1 year out of work and without a disability claim. Furthermore, in routine daily practice, filling out the FIS or FAL is only a part of the total examination and not the only way of report. However, the used procedure of the physicians (history and physical examination and filling out FIS and FAL) did not differ from the procedure in daily practice and therefore it can be concluded that using the FIS and FAL by social insurance physicians in this study result in insufficient reliability. To what extent this may have consequences in determination of disability claims in Social Insurance Medicine should be investigated.

REFERENCES

1. Roland M, Morris R. A study of natural history of back pain. Part 1: Development of a reliable and sensitive measure of disability in low-back pain. Spine 1983;8:141-144.
2. Deyo R. Comparative Validity of the Sickness Impact Profile and Shorter Scales for Functional Assessment in Low-Back Pain. Spine 1986;11:951-954.
3. Jensen MP, Strom SE, Turner J, Romano JM. Validity of the sickness impact profile Roland scale as a measure of dysfunction in chronic pain patients. Pain 1992;50:157-162.
4. Stratford PW, Binkley JM, Riddle DL. Development and initial validation of the back pain functional scale. Spine 2000;25:2095-2102.
5. Gommans IHB, Koes BW, van Tulder MW. Validiteit en responsiviteit Nederlandstalige Roland Disability Questionnaire; vragenlijst naar functionele status bij patiënten met lage rugpijn. Ned Tijdschr Fysioth 1997;107:28-33.
6. Beurskens AJ, de Vet HC, Koke AJ, van der Heijden GJ, Knipschild PG. Measuring the functional status of patients with low back pain. Spine 1995;20:1017-1028.
7. Bouter LM, Dongen MCIM van. Epidemiologisch onderzoek, opzet en interpretatie. Bohn Stafleu van Loghum, Houten/Diegem (4e druk), 2000.
8. Spanjer J. De inter- en intra-beoordelaarsbetrouwbaarheid van WAO-beoordelingen. Tijdschr Bedrijfs Verzekeringsgeneeskd 2001;8:235-241.
9. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37-46.
10. Landis RJ, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-174.
11. Altman DG. Practical statistics for medical research. Chapman & Hall, London, 1991.
12. Popping R. More agreeable software? Comput Stat Data Anal 1992;13:223-227.
13. Haas M. Statistical methodology for reliability studies. J Manipul Physiol Ther 1991;14:119-132.